KUOPION YLIOPISTON JULKAISUJA C. LUONNONTIETEET JA YMPÄRISTÖTIETEET 206
KUOPIO UNIVERSITY PUBLICATIONS C. NATURAL AND ENVIRONMENTAL SCIENCES 206

SAMULI-PETRUS KORHONEN

FLUFF-BALL, a Fuzzy Superposition and QSAR Technique

Towards an Automated Computational Detection of Biologically Active Compounds Using Multivariate Methods

Doctoral dissertation

To be presented by permission of the Faculty of Natural and Environmental Sciences
of the University of Kuopio for public examination in Auditorium L21,
Snellmania building, University of Kuopio,
on Saturday 27th January 2007, at 12 noon

Department of Biosciences / Chemistry
University of Kuopio

KUOPION YLIOPISTO
KUOPIO 2007

Distributor:
Kuopio University Library
P.O. Box 1627
FI-70211 KUOPIO
FINLAND
Tel. +358 17 163 430
Fax +358 17 163 410
http://www.uku.fi/kirjasto/julkaisutoiminta/julkmyyn.html

Series Editors:
Professor Pertti Pasanen, Ph.D.
Department of Environmental Sciences

Professor Jari Kaipio, Ph.D.
Department of Physics

Author's address:
Department of Biosciences / Chemistry
University of Kuopio
P.O. Box 1627
FI-70211 KUOPIO
FINLAND
Tel. +358 17 163 275
Fax. +358 17 163 259
E-mail: Samuli-Petrus.Korhonen@uku.fi

Supervisors:
Docent Mikael Peräkylä
Department of Biochemistry
University of Kuopio

Professor Reino Laatikainen
Department of Chemistry
University of Kuopio

Reviewers:
Dr. Mark Cronin
School of Pharmacy and Chemistry
Liverpool John Moores University
UK

Dr. Paul Lyne
Cancer Discovery, AstraZeneca
USA

Opponent:
Professor Mark S. Johnson
Department of Biochemistry and Pharmacy
Åbo Akademi

ISBN 978-951-27-0684-6
ISBN 978-951-27-0459-0 (PDF)
ISSN 1235-0486

Kopijyvä
Kuopio 2007
Finland

Korhonen, Samuli-Petrus. FLUFF-BALL, a Fuzzy Superposition and QSAR Technique – Towards an Automated Computational Detection of Biologically Active Compounds Using Multivariate Methods. Kuopio University Publications C. Natural and Environmental Sciences 206. 2007. 154 p.
ISBN 978-951-27-0684-6
ISBN 978-951-27-0459-0 (PDF)
ISSN 1235-0486

ABSTRACT

The presence of persistent organic pollutants in the global ecosystem is an unfortunate legacy of the explosive growth of the petrochemical industry in the 1960's. The EU has taken a proactive position and is implementing a comprehensive testing regimen for industrial chemicals; the so-called REACH legislation. Even with efficient *in vitro* and *in vivo* techniques this kind of extensive testing is a major undertaking because each molecule will have to be screened against hundreds of possible biological targets before a chemical can be declared safe. Thus, the benefits of computational, *in silico*, techniques become immediately obvious as biological activity could be predicted with computer using quantitative structure-activity relationship (QSAR) methods instead of arduous and expensive laboratory work. Unfortunately the data gained from QSAR are usually far from perfect, and in particular the generalisability and the automatisation of the structure response correlation (SRC) analysis have proven to be most elusive.

In this work a matching pair of superposition and QSAR techniques, the Flexible Ligand Unified Force Field – Boundless Adaptive Localized Ligand, is presented. FLUFF-BALL is designed to facilitate a rapid analysis of flexible molecule libraries with minimal user intervention. Primary design emphasis has been to maintain the computational simplicity necessary for fast screening while ensuring that the FLUFF-BALL remains easily tuneable allowing the user to import any and all available *a priori* information. In addition to FLUFF-BALL, MultiComponent Self-Organizing Regression, a novel PLS-type hybrid regression method is presented. The validation results clearly indicate that FLUFF-BALL is capable of generating robust predicting models for several different data sets and biological activities. In general the FLUFF-BALL generated results are comparable to those reported in literature. For highly congeneric systems the BALL was slightly inferior to the standard methods, but for a diverse xenoestrogen data set BALL met or exceeded the results of the standard 3D-QSAR method CoMFA. The results also indicate that the FLUFF superposition efficiently leverages available *a priori* information to dramatically improve the quality of superposition. For MCSOR the extensive validation runs clearly indicate that the MCSOR is a promising alternative and supplement to more established multivariate methods.

Universal Decimal Classification: 541.69, 57.014, 519.237, 504.064.2
National Library of Medicine Classification: QU 26.5, QV 26.5, QV 744, WA 671
Medical Subject Headings: quantitative structure-activity relationship; regression analysis; least-squares analysis; models, chemical; computer simulation; molecular structure; molecular conformation; environmental pollutants; xenobiotics

Vi veri universum vivus vici
- Faust

ACKNOWLEDGEMENTS

This study was carried out at the University of Kuopio, Department of Chemistry during the years 2001-2006 and I am grateful to the department for providing such excellent working facilities. Also the financial support from the graduate school in informational and structural biology was of vital importance. However, I would point out that the role of the graduate school is not limited to doling out the cash, but the school also provides and important chance to network, which is extremely important in a small country like Finland.

I wish to thank my main supervisor Docent Mikael Peräkylä for guidance, advice, and most importantly, moral support during the long and demanding research work. I can with confidence say that as much as I benefited from our scientific discussions, the enjoyment I got from our meandering and rather informal discussions also greatly aided the formation of this thesis.
I would also like to thank my other supervisor Prof. Reino Laatikainen for his enthusiasm and endless ideas. I also appreciate those many discussions I had with him about many things including, but not limited to, science. Last, but by no means least, I am grateful for Docent Kari Tuppurainen who mentored me in the finer points of QSAR and statistical analysis. His patient help and guidance was invaluable as I took my first faltering steps on the long road to the skills needed of an independent researcher.

Furthermore, I express my gratitude to the reviewers, Dr. Mark Cronin and Dr. Paul Lyne, for spending their time in reading this thesis and for the valuable feed-back they gave. I am also grateful for Ewen McDonald and Niko Jukarainen who had the patience to proof-read the English in this thesis.

I am eternally grateful to parents Hannu and Ranja and for all my friends who provided a much needed life-line away from academia. They allowed me to forget my thesis, as much such a thing is possible, during my free time. They put things in proper perspective by reminding me that life is so much more than work and thus they enabled me to put aside all the stresses and frustrations of research and simply enjoy life.

Finally, I wish to thank all of my co-workers at the Department of Chemistry for their supportive and friendly atmosphere. It has, and will be, a pleasure to work with you.

Kuopio, January 2007

Samuli-Petrus Korhonen

ABBREVIATIONS

ANN	artificial neural network
BALL	boundless adaptive localized ligand
CoMFA	comparative molecular field analysis
CV	cross-validation
E_2	17b-estradiol (estradiol)
EDKB	endocrine diruptor knowledge base
FLUFF	flexible ligand unified force-field
kNN	k-nearest neighour
LMO	leave-many-out (a form of cross-validation)
LOO	leave-one-out (a form of cross validation)
MCSOR	multi-component self-organizing regression
MLR	multiple linear regression
NPC	number of principal components
PCA	principal component analysis
PCR	principal component regression
PLS	partial least squares
PRESS	predictive residual sum of squares
Pr-R^2	predictive correlation coefficient
Q^2	cross-validated correlation coefficient
QSAR	quantitative structure activity relationship
QSPR	quantitative structure property relationship
RR	ridge regression
SAR	structure activity relationship
SDEP	standard error of prediction
SOM	self-organizing map
SOMFA	self-organizing molecular field analysis
SOR	self-organizing regression
S_{press}	cross-validated standard error of prediction
SRC	structure response correlations
VDW	van der Waals

LIST OF ORIGINAL PUBLICATIONS

This doctoral thesis is based on the following original publications:

I. Korhonen, S.-P.; Tuppurainen, K.; Laatikainen, R.; Peräkylä, M. (2003). FLUFF-BALL, A Template-Based Grid-Independent Superposition and QSAR Technique: Validation Using a Benchmark Steroid Data Set, *Journal of Chemical Information and Computer Sciences* 43, 1780-1793

II. Korhonen, S.-P.; Tuppurainen, K.; Laatikainen, R.; Peräkylä, M. (2005). Comparing the Performance of FLUFF-BALL to SEAL-CoMFA with a Large Diverse Estrogen Data Set: From Relevant Superpositions to Solid Predictions, *Journal of Chemical Information and Modeling* 45, 1878-1883

III. Korhonen, S.-P.; Tuppurainen, K.; Laatikainen, R.; Peräkylä, M. (2005). Improving the performance of SOMFA by use of standard multivariate methods, *SAR and QSAR in Environmental Research* 16, 567-579

IV. Korhonen, S.-P.; Tuppurainen, K.; Asikainen, A.; Laatikainen, R.; Peräkylä, M. (2005). SOMFA on large diverse xenoestrogen dataset: The effect of superposition algorithms and external regression tools, *QSAR and Combinatorial Science*, *In Press*

V. Tuppurainen, K.; Korhonen, S.-P.; Ruuskanen, J. (2006) Performance of Multi Component Self-Organizing Regression (MCSOR) in QSAR, QSPR, and Multivariate Calibration: Comparison with Partial Least-Squares (PLS) and Validation with Large External Data Sets, *SAR and QSAR in Environmental Research* 17, 549-561

CONTENTS

1. A BRIEF HISTORY OF STRUCTURE RESPONSE CORRELATIONS

When faced with the challenging task of screening large libraries of molecules for biological activity, be it for drug discovery or for the identification of possibly hazardous molecules, the benefits of computational, the so-called *in silico*, prediction of biological activity become immediately obvious. Instead of arduous and expensive laboratory work to measure the biological activity, it could be predicted with a suitable *structure response correlation* (SRC) technique which only requires computing capacity[1]. In principle, these techniques could be used as a replacement for animal testing, but even if this ultimate goal proves to be a too tall order, they can be used to streamline the synthesis and screening of new drugs which will result in considerable savings for the pharmaceutical industry[1,2].

The field of environmental chemistry, in particular, would greatly benefit from a reliable *in silico* tool for the prediction of the biological activity as the widespread use of synthetic chemicals has led to a veritable explosion in the number of xenobiotic chemicals present in the ecosystem. One of the major problems with xenobiotics is that, for many compounds, due to their unnatural structure, no biological degradation pathway has evolved and these chemicals will inevitably start to accumulate in the ecosystem. As the concentration of a chemical increases, the likelihood of adverse health or environmental effects also increases which is due to the fact that for high concentration even a low intrinsic biological activity would be sufficient to cause a marked response. Even though the testing of a limited number of chemicals against a single biological target is rather simple, the testing of an extensive molecular library with *in vivo* or *in vitro* techniques is virtually impossible. Even more so, as there are hundreds of possible biological targets which would have to be screened against before a chemical can be declared safe. Thus the SRC techniques have become particularly interesting as the EU is currently implementing a comprehensive testing regimen for industrial chemicals, the so-called REACH legislation[3], and a vast number of the chemicals in use today must also be tested, the benefits of computational screening of hazardous chemicals by using SRC would be tremendous.

Unfortunately, the activity data gained from a SRC analysis are usually far from perfect and despite intensive efforts there is no universal solution for the structure-based prediction of the biological activity for a diverse set of compounds. Also, the automatisation of the SRC analysis has also proven to be most elusive and considerable amount of human intervention is required. More worryingly, the building of a SRC model often requires chemical intuition, which means that there is still a great deal of subjectivity in the model and thus the reliability of the model must be rigorously tested using approved statistical methods[4-7].

Next a short introduction outlining the historical roots and the major methodological developments in SRC is presented. This short introduction is by no means a comprehensive tutorial to the early SRC techniques and many methodological advances have been omitted. For a more comprehensive view the reader is referred to reviews by Kubinyi[2] and Rekker[8] and references therein.

The roots of structure-response correlation analyses can be traced back to the 1860s when some observations of a correlation between molecular structure and biological activity were reported. One could argue that as these observations were anecdotal and no systematic effort was made to generalise the observations, they did not fulfill the criteria for a true SRC analysis[2,8] Those observations however, paved the way for the principle behind all structure response correlation analyses: *"The response of a system to a chemical compound depends only on its structure."* For this simple principle one can further deduce that if the response (Φ) is dependent only on the structure of a compound (*C*), there must exist a function (*f*) describing the correlation between the structure and the response (eq. 1) This principle was first formulated by A. C. Brown and T. Frazer in 1868 and it has become the cornerstone of the SRC analysis[2,9].

$$\Phi = f\left(C\right) \tag{1}$$

The basic principle only infers that there is a relation, but it does not indicate anything about the exact nature of interdependence between the response and structure. Furthermore, it does not specify how the structure of a compound should be described. Yet, from this early formulation we can already separate the parts of a modern SRC analysis. On the left side we have the response elicited by the compound (Φ), and on the right side we have function correlating the structure with activity (*f*). In order to keep the correlating function as simple as possible one should use a suitable function (*d*) to derive a mathematical representation of the structure (*D*), called a *descriptor*, and feed it to the correlating function (eq. 2). The descriptor can be a single number, but it can also be a vector, or a matrix of a lattice of numbers describing the physicochemical properties of the compound, hence the name descriptor.

$$\begin{aligned} D &= d\left(C\right) \\ \Phi &= f\left(D\right) \end{aligned} \tag{2}$$

The Meyer-Overton model of narcosis[2] (eq. 3), presented at the turn of the 20th century, is arguably the first true SRC model as it explicitly links the narcotic power (*N*) of alcohols, ethers and amides with the logarithm of their olive-oil-water partition coefficient (Log P). Of course the descriptor variable used is in itself a property but it is directly dependent on the structure of the compound and thus it can be interpreted as a SRC model.

$$\log\left(\frac{1}{N}\right) = 0.94\log P + 0.87 \tag{3}$$

Even though the earliest formulation of the SRC equation correlated the properties of a compound directly with its response, it was soon discovered that it is very difficult to formulate the correlation function (*f*) as it can be very complex.

On the other hand, if the response is known for a set of structurally similar molecules, it is relatively easy to correlate the change in the response ($\Delta\Phi$) with the change in the structure of a compound (ΔC), as shown in equation 4. The primary benefit of this approach is the fact that the function (f^{Δ}), which correlates a small change in structure to a small change in response ($\Delta\Phi$), is usually very simple and thus easy to formulate.

$$\Delta\Phi = f^{\Delta}\left(\Delta C\right) \tag{4}$$

The next major contribution to the development of SRC analysis came in the 1930's as L. P. Hammett studied the ionisation constants of substituted benzoic acids. He postulated that the differences in the reaction rate constants (k, eq. 5) between the unsubstituted compound (H) and the compound containing a substituent (X) linearly depend on the reaction specific constant (ρ) and the Hammett constant (σ, eq. 6). If a compound has several substituents their effects are assumed to be fully additive. Thus, Hammett moved from the direct correlation of structure and response to the incremental model where the change in the structure is correlated with a change in response. From the start these physico-chemical equations were used to derive many formulations of biological and biochemical Hammett equations. However, due to the simplicity of the linear model and a single variable descriptor, the results were usually poor, even though there were some successful models[2].

$$\log k_X - \log k_H = \rho\sigma \tag{5}$$

$$\log K_X - \log K_H = \sigma \tag{6}$$

In the 1950's Taft extended the Hammett equation, which originally only considered the electronic effects of the substituents, by defining a new term E_s (eq. 7), with which one can also take into account the steric hindrances caused by bulky substituents. The new term added only a modest amount of predictive power and by 1960 it was clear that without major methodological improvements the SRC analysis was at a dead-end[2,8].

$$E_s = \frac{\log k_X}{\log k_H} \tag{7}$$

In the mid-1960s C. Hansch formulated an equation for predicting the water-octanol partition co-efficients (eq. 8). An extension of this equation was proposed by T. Fujita who suggested that the predictive power of the original model could be increased by combining several descriptors into a one equation (eq. 9)[2]. In 1964 Hansch and Fujita published a paper in the Journal of the American Chemical Society entitled "*The ρ-σ-π analysis. A method for the correlation of biological activity and chemical structure*", which outlined the method now known as classical QSAR[2,10].

$$\pi_x = \frac{\log P_X}{\log P_H} \tag{8}$$

$$\log\left(\frac{1}{C}\right) = k_1\pi + k_2\sigma + ... + k_m \tag{9}$$

Almost at the same time as Hansch and Fujita published their ground-breaking paper, another pair of researchers, Free and Wilson, had formulated an alternative way of performing SRC analysis (eq. 10). In their approach the compound is described by a long binary vector (*I*) where each binary digit (bit) indicates the presence or absence of a certain feature at a specific location. This vector is multiplied by a vector of regression coefficients, called enhancement factors (*F*), in order to generate the SRC model. The problem with Free-Wilson analysis is that as the number of substituents and the number of substitution sites grows, the length of the indicator vectors grows very rapidly due to the so-called combinatorial explosion. For example, if a new substituent is introduced to the Free-Wilson model, the length of the indicator variable grows by the number of unique substitution sites in the backbone. Despite these limitations, Free-Wilson analysis has proven to be quite a useful tool in combinatorial chemistry where, due to the limitations of synthesis capability, the length of *I* is limited and also the chemistry is usually limited to a single backbone, meaning that the molecules form a congeneric set[2,9].

$$R = \sum F_i I_{ni} + k \tag{10}$$

It is also possible to combine the Hansch and Free-Wilson models to form mixed models where the ρ-σ-π parameters describe the large changes in the molecule structure, while at the same time, the Free-Wilson analysis describes the exact changes in substitution. In later years the original ρ-σ-π analysis was extended by C. Hansch and others by the inclusion of the second powers of the terms to generate a parabolic model (eq. 11), or cross-terms (eq. 12), in order to compensate for the non-linearity of the structure response correlation. Additionally, many new descriptors and alternative formulations of the classical QSAR equation were made, and the whole field of SRC begun a period of rapid development[2,9].

$$\log\left(\frac{1}{C}\right) = k_1\pi + k_1'\pi^2 + k_2\sigma + k_2'\sigma^2 + k_3E_s + k_3'E_s^2 + ... \tag{11}$$

$$\log\left(\frac{1}{C}\right) = k_1\pi + k_2\sigma + k_3\sigma\pi ... \tag{12}$$

The major methodological breakthrough was the introduction of 2D descriptors based on the topological analysis of the molecular structure. In a few short years a plethora of these new descriptors were formulated[9], and as new SRC descriptors were developed, the scope of the different types of problems the SRC was applied to, also expanded[2,9].

With the new classes of problems the number of SRC techniques increased rapidly and soon the SRC became a family of techniques. A summary of the main branches of SRC analysis is presented in Table 1. Though several other types of SRC analysis have been introduced in the literature, such as quantitative spectrometric data-activity relationship (QSDAR)[11], or quantitative structure-biodegradability relationship (QSBR)[12], but they have not been widely adopted and have therefore been omitted from this summary.

Table 1. The different types of SRC analyses. The types of source data are indicated in the columns and the response types are listed on the rows.

	STRUCTURE	**PROPERTY**
ACTIVITY	(Quantitative) Structure-Activity Relationships *(Q)SAR*	(Quantitative) Property-Activity Relationships *(Q)PAR*
PROPERTY	(Quantitative) Structure-Property Relationships *(Q)SPR*	
TOXICITY	(Quantitative) Structure-Toxicity Relationships *(Q)STR*	
RETENTION TIME	(Quantitative) Structure-Retention Relationships *(Q)SRR*	

Even though the classical QSAR was quite successful, there were pathological cases, usually involving stereospecificity, indicating that the SRC descriptors based on the 2D information could not fully encompass the complexity of a chemical structure. Also, many classical SRC analysis methods could only process congeneric sets, i. e. molecule sets with common backbone structure, which limited their usefulness[2,9,13]. So, in 1979 R. D. Cramer proposed a new SRC analysis paradigm called *dynamic lattice-oriented molecular modelling system* (DYLOMMS) which could circumvent both problems. In this approach the molecule is embedded in a three dimensional orthogonal isotropic grid. Then the descriptor is evaluated by computing the values of molecular property fields, such as steric repulsion or attraction and electrostatic potential, at the vertices of the lattice[9]. Unfortunately the statistical analysis methods of the day were not up to the task and the DYLOMMS approach was not generally accepted, mainly due to unremarkable results. Regardless of such an inauspicious start, the next generation of the grid-based SRC analysis methodologies emerged in the 1980's, when the Partial Least-Squares (PLS) regression methodology was applied to SRC problems. These techniques addressed some of the major deficiencies inherent in classical QSAR techniques, and they were soon widely adopted. The most notable ones of these new methods were Comparative Molecular Field Analysis (CoMFA) by Cramer[14] and GRID by Goodford[15], of which the CoMFA gained wider acceptance[9].

The advances in computer hardware, which rapidly increased the availability of sufficiently powerful computers, lead to a veritable gold rush to employ 3D SRC as an aid in medicinal chemistry and several other fields[16]. Unfortunately in their haste many practitioners forgot the limitations of this new technique and a horde of appallingly poor models were published. In particular, a common problem was the lack of sufficient validation and therefore the true predictive power of many models was rather poor. This also led to an over-interpretation of the

models which in turn often led to erroneous conclusions[4,5]. So widespread were these problems that a mere decade later there was ample evidence that 3D SRC is not the panacea which it was hoped to be[5,17]. On the other hand, it has also become clear that by careful model building and, more importantly, by stringent validation of the model it is possible to create highly predictive 3D SRC models that can prove to be invaluable tools in trying to rationalise the observed responses[1].

Although the grid-based SRC techniques are generally considered to be the most effective means of predicting biological activity, they usually require an accurate superposition of structures, which has proven to be a major bottleneck[4,6,7,18,19]. The alignment procedure usually requires considerable human intervention and is generally regarded to be the most arduous and time-consuming phase of the grid-based SRC analysis. The requirement of accurate superposition also severely limits the efficiency of these techniques when dealing with large and diverse molecule sets[20]. Therefore a fully automated computational "sieve", capable of rapidly browsing through vast molecular libraries, and eliminating the non-active compounds, would be very useful in many applications. For the above reasons, considerable effort has been directed into the automation of the superposition process but unfortunately, a definite solution has not been found. Furthermore, it seems that no universally applicable automated solution will be found in the near future and the multitude of different algorithms, each of which being a partial solution, continues to exist[18,21].

It should be emphasised that the words "system" and "response" used in the definition of structure response correlations are purposefully vague. This means that the same basic principles underlying the SRC methodology can be applied to a great variety of systems responding in almost any conceivable manner to a compound. The SRC can be used to model something as simple as few chemicals in a test tube or a whole organism. A few models have even tried to model a complete ecosystem. Naturally the complexity of the model and the amount of the experimental source data required for a reliable model also rapidly increases as the complexity of the modelled system increases. Regardless of the problems inherent in the prediction of complex systems, the emphasis of the SRC analysis is slowly shifting from the prediction of simple receptor binding to a more holistic approach including the modelling of more complex systems. One of the rising areas of application for SRC is the prediction of so-called ADME/Tox properties of new potential drugs. The first part of the abbreviation refers to the words Absorption, Distribution, Metabolism and Excretion (ADME) which, when linked with Toxicity, forms ADME/Tox. Many models have been reported in the literature[22-59] but it is clear from the results that despite the many success stories, methodological advances are required before the *in silico* ADME/Tox is sufficiently reliable to be routinely used to assess the viability of the potential drugs before *in vivo* experiments[45,60,61].

As a summary, one could say that in the 75 years since Hammett equation, the SRC has become an important tool for many disciplines, more so as it works well as a complementary tool expanding and refining the information available for experimental work. As is typical with advances in science, the SRC never was quite as powerful as the optimists hoped for, but at the same time it was not as useless as the pessimists feared, but all in all it has proven to be an invaluable tool. Therefore, one can predict with confidence that the SRC will play an important

role in the future. More so as the SRC is far from maturity and the rapid methodological development along with the ever increasing computing capacity will undoubtedly increase the flexibility, power and stability of SRC techniques in the future[1].

2. THE PHASES OF A SRC ANALYSIS

A simplified flowchart indicating the phases of a typical structure response correlation analysis is presented in Figure 1. It starts with a set of molecules with limited structural variability and known responses, which is used to derive a corresponding set of descriptors. These data are fed into a statistical analysis method in order to derive an estimate of the correlation function ($f^{\triangle}$) whose fitness is then evaluated in order to estimate the predictive power of the SRC model. Some of the phases, such as descriptor post-processing or visualisation, may be omitted for some SRC analysis methodologies, but usually nearly all of the phases are needed to build a SRC model. In the following sections an overview of each phase will be presented including a few fundamental references which enable the reader to gain more in-depth information.

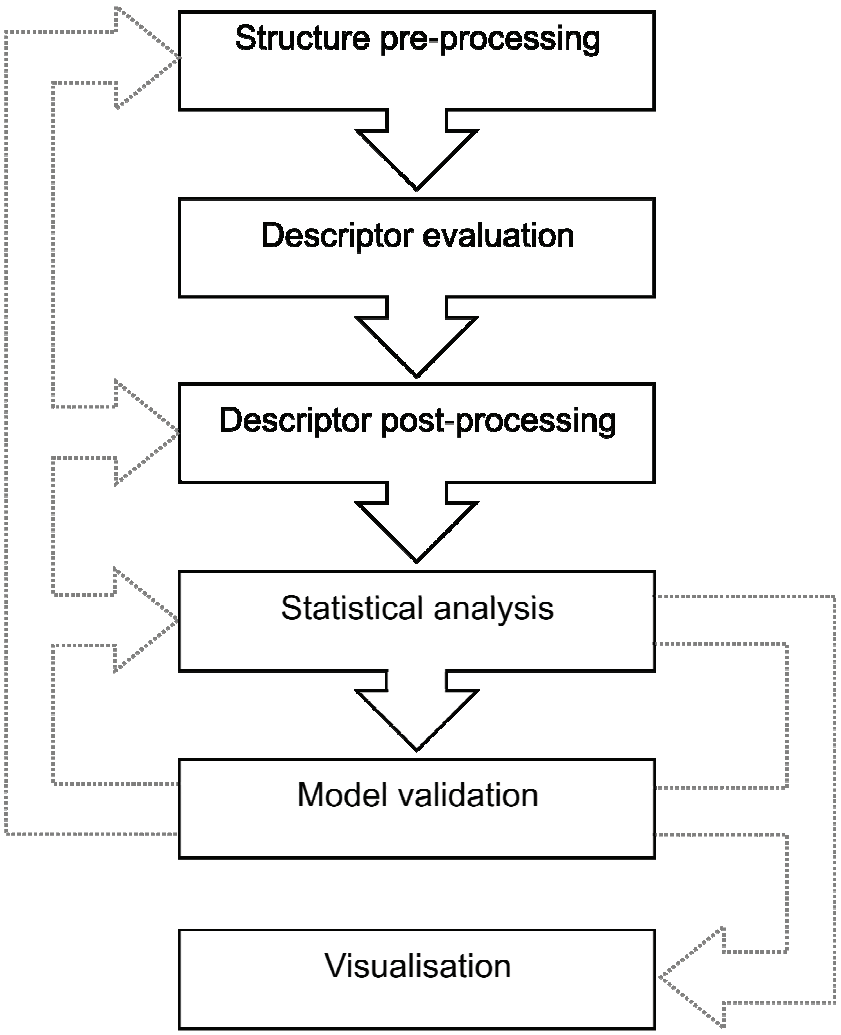


Figure 1. Flowchart of the phases of a typical structure response correlation (SRC) analysis.

2.1 Structure pre-processing

The first step in any structure based SRC analysis is the structure generation by some molecular modelling software. For those SRC techniques that require the 3D structure one must also perform a geometry optimisation in order to obtain a reasonable starting conformation. Also, many techniques need accurate partial charges and their evaluation usually requires the 3D structure, so in most cases one is forced to compute the geometry optimised 3D structures regardless whether the SRC technique requires it or not. Fortunately the structure generation and optimisation are nowadays standard operations and by using a semi-empirical quantum mechanical computation the charges can be evaluated very quickly and reliably. It is also customary to optimise to geometry of the structure. One can also use more sophisticated *ab initio* quantum mechanical computations, but usually this does not improve the accuracy of the SRC model as the errors stemming from other sources are greater than the ones caused by the inaccuracies of the optimisation method. In particular, it should be emphasised that the optimisation is usually performed using a gas phase, *in vacuo*, model or an implicit solvent, but in the experimental systems the compounds are fully solvated and therefore a massive optimisation is not worth the effort as the differences between the optimisation techniques are smaller than the effects caused by the solvation.

2.1.1 Structure Alignment

Some SRC analysis methodologies, especially the so-called grid-based 3D techniques, require an alignment of the molecular structures. The basic design of all the alignment techniques is the same: First, one must have a metric which measures the similarity of the molecules superposed. In other words the metric indicates the goodness of a particular alignment. Secondly one must also have a transformation function which generates new alignments. Finally, there must also be an optimisation method to guide the transformation towards optimal alignment. The actual alignment is a process in which the similarity, as indicated by the metric, is optimised by using an appropriate mathematical method in conjunction with a transformer function.

In the context of structural alignment the transformation function corresponds to the degrees of freedom available in the system. If the alignment is rigid, the molecule can only change its orientation and position, but its conformation does not change. So basically the system has 6 degrees of freedom, namely a translational and a rotational degree of freedom for each axis. Then one can have a semi-flexible system where a part of the structure is flexible while the rest is rigid. A most typical case of this kind of system consists from a rigid backbone and flexible substituents. To prevent the flexible parts from adopting energetically extremely unfavourable conformations a set of constraints, often in the form of a molecular mechanics force field, must be applied. Even in the semi-flexible system the number of degrees of freedom drastically increases to $3N_f+6$ where N_f is the number of flexible atoms. In a fully flexible system all atoms can move freely. However, in order to keep the molecular structure intact the constraints become even more important than in the case of the semi-flexible algorithms. The maximum number of degrees of freedom for a fully flexible system is 3N, where the N is the number of atoms. Naturally these maximum degrees of freedom are usually not available as the constraints allow only a small subset of possible atom positions. Also an average molecule often has rather

rigid parts, like an aromatic ring, which further decreases the effective atom count. So in many cases one can reduce the molecule into a set of rigid fragments connected by flexible parts. Nevertheless, all these short-cuts, even though they considerably increase the performance of the alignment, are in essence heuristic systems whose performance can not be guaranteed for all molecules[62]

When doing the alignment, often also referred as superposition, one should first check whether there is an unambiguous binding site into which all molecules will fit and is the structure of the binding site on receptor or enzyme known. If that is the case then it would make sense to utilise the additional information provided by the receptor structure, the more so as it has been demonstrated that an alignment generated using constraints derived from the receptor model are in many cases superior to the standard ligand based alignments[63,64]. These receptor structure-based alignment techniques are called ligand docking techniques[65-68] as they "dock" the ligand molecule into a cavity on the receptor structure. They can be used to predict binding orientation[69,70] and even binding affinity[15,65] of the ligands. The critical point in ligand docking is the scoring function which evaluates the fitness of the found binding orientation and output the so-called docking score. Unfortunately a reliable and universally applicable scoring function has not been found which limits the usefulness of these techniques[67]. The detailed description of these techniques does not fall into the scope of this work and for further details on ligand docking the reader is referred to recent reviews by Krovat et al[67] and Taylor et al[68].

The so-called point-based algorithms represent the simplest form of a similarity metric. They simply measure the distances between a set of points, usually comprising of pairs of atoms called *anchor points*, and use these distances to measure the similarity of the molecules. For congeneric molecules it is usually easy to decide which atoms should be aligned but as the diversity of the molecules increases it becomes increasingly difficult to generate these points. The anchor points are usually assigned manually and thus the alignment requires a considerable amount of human intervention. Therefore algorithms for automatic detection of anchor pairs have been proposed, but in general the performance of such techniques has remained modest[71]. Instead of a physical molecule one can also use an abstract structure called a *pharmacophore* which indicates the molecular features essential for biological activity. The pharmacophore is composed from pseudoatoms defining the necessary steric and electrostatic properties as well as hydrogen bond properties.

A more complex, and also much more successful, set of distance metrics is the field fitting or property-based techniques which try to measure the similarity of the two molecules using "fields" generated from molecular properties using suitable functions. These property-based algorithms offer a wide choice of descriptors, which include molecular shape and volume, electron density, charge distribution and many more. Yet, all these techniques generate a number (or a set of numbers), usually called *similarity indicies*, which describe the degree of similarity between the fields generated from the structures which are being superposed. At the moment these techniques are the methods of choice for alignment of molecules for SRC analysis and they have also found use as SRC descriptors[62].

In recent years a new set of metrics, which use the shape of the molecular surface in conjunction with local properties to find the optimal alignment, have been developed[72-76]. They are intriguing and could be much more efficient than the field fitting metrics, but at the same time they are also rather new techniques and more experience about their behaviour is needed before any conclusions can be drawn about their eventual performance.

Unfortunately, the similarity metric usually has numerous local minima and thus it has proven very difficult to find an optimisation technique that could reliably find the global optimum regardless of the initial position of the molecules. In many cases a normal optimisation technique is used and it is up the user to ensure that the initial guess provided to the alignment algorithm is near the global optimum or otherwise the user will end up with an alignment that corresponds to a local minima. Some superposition algorithms use minimum elimination or stochastic optimisation techniques, such as Monte Carlo or poling, to escape the local minimum in order to find the global minimum[18,77]. All in all the problem of finding the global optimum of a complex function, such as the similarity metric, is still a mathematically unsolved problem and will very likely remain as such for the foreseeable future.

Despite intense efforts devoted into the development of alignment techniques it seems that the automatisation and universal applicability will remain unattainable for the foreseeable future and semi-automated solutions such as QXP[78], SEAL[79] and many others[18,21,72-76,80-93], represent the best available techniques. For an extensive bibliography on the available superposition algorithms, with a particular emphasis on the different metrics, the reader is referred to the papers by Melani et al[18], Lemmen et al[21] and to references therein.

2.2 Descriptor categories and their dimensionalities

When the molecules have been generated and, if necessary, aligned, they are usually loaded into a program which does the actual evaluation of the descriptor. There are several programs available, such as Sybyl[94], ALMOND[20], Quasar[95], DRAGON[96], CODESSA[97], TAM[98], TOP[99] and many others[9]. Some of them are capable of evaluating several different types of descriptors whereas others are dedicated to only one descriptor implementation. These programs have become so sophisticated that it is usually very simple to do the actual evaluation.

On the other hand, it is often a more difficult problem to decide which descriptors to use as there are literally hundreds of different formulations. The greatest problem in selecting a descriptor is that there are considerable case-per-case differences in their performance and therefore it is very difficult to say with any confidence, that a certain descriptor will be optimal for this particular set of molecules[5,100]. To summarise one could say that the current state of SRC descriptors is reminiscent of the situation with alignment algorithms as there is a overabundance of different descriptors, each of which is good at some things and poor at others. It would seem that this ambiguity will persist as it is likely that a universally applicable QSAR technique will not to be found in the near future, despite intensive efforts[5].

One should also bear in mind that even though SRC descriptors may seem to be radically different they, none the less, exhibit strong mutual correlations and therefore only a modest increase in predictive ability is achieved by combining different descriptors and thus this kind of quorum thinking can not be used to circumvent the descriptor selection problem. On the other hand, there are clear indications that by using the so-called consensus methods, it is possible to partially compensate for the inherent bias in the descriptor. The idea of the consensus technique is to generate a large number of independent SRC models from a same molecule set and then use the individual predictions as second order descriptors to evaluate the final prediction[101-104]. One can use a simple weighted average of the individual predictions or one can also use more sophisticated statistical analysis tools to detect and discard outliers. In any case, it is unlikely that for a compound the majority of the SRC techniques would err in the same direction. Therefore the use of several independent techniques will tend to even out the individual errors and one can also use the distribution of the predicted values as an indicator of the reliability of the prediction.

In the following subsection a brief overview of the available SRC descriptors is given. Due to the vast number of different methodologies available this short chapter can not be a definitive reference but it rather tries to give broad outlines of the different techniques and the mathematical principles which they are based upon. For a more in-depth discussion on the subject of descriptors the reader is referred to the *Handbook of Molecular Descriptors* by Todeschini and Consonni[9] which includes an extensive bibliography of about 3000 references.

2.2.1 0D to 2D descriptors

Zero-dimentional (0D) descriptors utilise the atomic or molecular properties and are therefore independent of the overall molecular connectivity. They include a great variety of descriptors such as molecular mass or refractivity, element count, element quotient, and many others[9]. Some practitioners point out that descriptor values should be derived using mathematical and logical procedures and therefore the property-based descriptors should be excluded from the SRC[9]. If one accepts molecular properties as descriptors, they should formally be considered dimensionless, but they are often included among the 0D descriptors. The property-based descriptors include many kinds of empirical parameters including among others the Hammett, Taft and Hansch constants discussed earlier. One of the more complex property-based techniques is Comparative Spectra analysis (CoSA)[11,105-107] in which an experimental spectra is transformed into a vector which in turn is used to generate the descriptor. In principle this technique could use any spectral information, but in practice the 1D NMR spectra has been used almost exclusively. An extension of the CoSA, called Comparative Structural Connectivity Spectral Analysis (CoSCoSA)[108], has also been proposed. In this technique different NMR experiments, which are sensitive to the configuration and conformation, are used to generate the descriptor.

The 1D descriptors are based on the local or fragment connectivity. They include several types of fragment counts, branch indices, ring counts and molecular fingerprint descriptors[9]. The HQSAR methodology is an interesting combination of a compressed molecular fingerprint, a normal 1D descriptor and 2D topological analysis. It uses the 1D descriptor to build the SRC model but at the same time it also utilises 2D information as it contains a phase where the set of fragments used to compute the molecular fingerprint is automatically generated. Despite of its relative simplicity the HQSAR has proven to be a very effective form of SRC analysis[7,109-127].

The 2D descriptors should be called topological descriptors or "graph invariants" as they are evaluated using the so-called molecular graph which is constructed by replacing atoms with vertices and bonds with edges. Usually the topological indices are computed from the so-called *hydrogen depleted molecular graph* in which hydrogen atoms are omitted while building the graph. One should note that the molecular graph is a topological construct and does not take the 2D or 3D structure into account. Therefore the three molecular graphs presented in Figure 2 (B-D), even though they look different, are all equally valid topological descriptors for 2-methyl-pentane (A).

A large class of 2D descriptors, called topostructural indices, rely solely on the molecular graph and therefore take into account only the topology of the molecular graph and discard the chemical information available about the underlying compound. These descriptors are derived using purely topological and graph theoretical principles. Popular topostructural 2D descriptors include the Wiener index[128-141], Zagreb index[142-146], the Randic connectivity index[147-151] and Balaban index[138,152-158]. For example, one can compute the Wiener index for 2-methyl-pentane using equation 13 and a distance matrix (eq. 14) which indicates how many edges (bonds) one must cross in order to get from atom *i* to atom *j*. After rather simple arithmetic (eq. 15) one gets the Wiener index value 32 for this molecule.

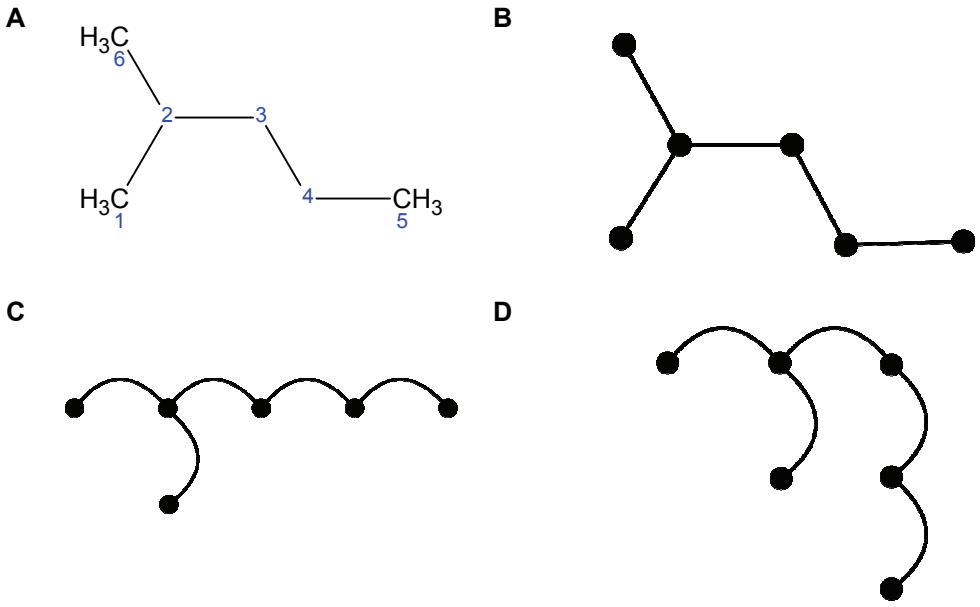


Figure 2. 2D structure of a molecule (A) and three equivalent hydrogen depleted molecular graphs (B-D)

$$W = \frac{1}{2}\sum_i \sum_j d_{ij} \tag{13}$$

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 2 \\ 1 & 0 & 1 & 2 & 3 & 1 \\ 2 & 1 & 0 & 1 & 2 & 2 \\ 3 & 2 & 1 & 0 & 1 & 3 \\ 4 & 3 & 2 & 1 & 0 & 4 \\ 2 & 1 & 2 & 3 & 4 & 0 \end{bmatrix} \end{array} \tag{14}$$

$$\begin{aligned} W = \frac{1}{2}\big(&(0+1+2+3+4+2)+(1+0+1+2+3+1)+\ldots \\ &(2+1+0+1+2+2)+(3+2+1+0+1+3)+\ldots \\ &(4+3+2+1+0+4)+(2+1+2+3+4+0)\big) \\ W = \frac{1}{2}&(12+8+8+10+14+12) = \frac{1}{2}(64) = 32 \end{aligned} \tag{15}$$

Because the topostructural descriptors overlook a considerable amount of information available in the 2D structure a set of topochemical descriptors have been formulated which are sensitive to the topology of the molecular graph and to the chemical information it contains. Some topochemical descriptors are extensions of topostructural indices[134,159-162], but there are also completely novel indices such as Kier-Hall[163-167], Burden eigenvalues and many others[9]. There also exists a set of 3D topological descriptors which are extensions of the 2D topostructural indices. These 3D indices use the 3D Cartesian inter-atomic distances instead of the discrete topological distances and are therefore sensitive to the 3D conformation of the molecule [9].

One important class of 2D descriptors are the electorotopological (E-State) descriptors which describe the electronic state of the molecule. Many formulations and even more numerous applications have been reported in the literature[163,168-188]. While the reported performance has varied greatly, it is clear that in many cases the E-state indices can create models which are comparable to those derived from 3D descriptors[163,177,179,187]. One of the more popular E-state descriptors is the Molecular Electronegativity Distance Vector (MEDV) which has been successfully used to model many systems[12,189-194].

For a more detailed discussion on the available 0D to 2D descriptors the reader is referred to the *Handbook of Molecular Descriptors* by Todeschini and Consonni[9].

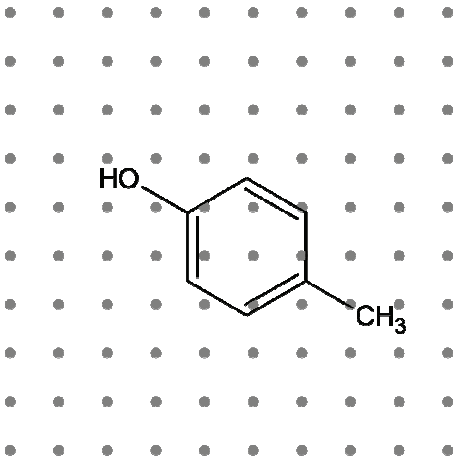
2.2.2 3D descriptors

As stated earlier, the usefulness of the 3D SRC techniques is often limited by the need for an accurate superposition of structures[18,62]. This is primarily due to the fact that there is no universally applicable automatic superposition method. Even the currently available semi-automated solutions usually require considerable human intervention when dealing with large and diverse libraries of molecules. Despite these problems the 3D SRC, and the grid-based techniques in particular, have been widely utilised.

The conceptually simplest 3D SRC analyses are the grid-based techniques, which embed the molecules in a three-dimensional grid and derive the descriptor by evaluating the values of descriptor functions at the vertices of the grid. A schematic 2D representation of a grid-based SRC (Figure 3A) is used instead of a real 3D as it is easier to represent on paper. The ligand surrounded by the grid, which is used to derive descriptor matrices (Figure 3B) for van der Waals and electrostatic interactions. Also many alternate descriptor fields, such as HINT[195] which describes the hydrophobic interactions, have been proposed[9]. This enables one to easily extend the grid-based SRC descriptors from the default steric and electrostatic to almost any conceivable property. For further details see a paper by Kellogg[196] and the Handbook of Molecular Descriptors[9]. The separate matrices are combined and transformed into a single vector which is then fed to the statistical analysis methods in order to derive the actual model. Yet, one can still easily connect the variables to the grid points, which has advantages when one visualises the result (see page 52).

In addition to the problems caused by the superposition, 3D SRC techniques that depend on a global grid are also susceptible to errors rising from translation and rotation of structures. One can quite easily see that if one were to rotate or translate a molecule in the grid the way in which its atoms intersect with the grid, vertices also change. Unfortunately this can lead to changes in the predictive power of the model and therefore the overall position and orientation of the molecules becomes critical. One can reduce this effect by increasing the resolution of the grid but this leads to an exponential growth in the number of variables creating problems in the statistical analysis. Thus the resolution of the grid is always a balancing act between the number of variables and rotational and translational sensitivity[20,197-199].

A



B

$$\begin{bmatrix} s_{1,1} & \cdots & s_{1,10} \\ \vdots & \ddots & \vdots \\ s_{1,10} & \cdots & s_{10,10} \end{bmatrix} \begin{bmatrix} e_{1,1} & \cdots & e_{1,10} \\ \vdots & \ddots & \vdots \\ e_{1,10} & \cdots & e_{10,10} \end{bmatrix} \Rightarrow \left[s_{1,1} \ldots s_{10,10}, e_{1,1} \ldots e_{10,10} \right]$$

Figure 3. A Schematic representation of a grid-based SRC technique (A) and the resulting matrices (B)

Grid-based techniques include CoMFA[14], GRID[15] and SOMFA[200]. Of these three the CoMFA has been particularly popular and at the end of 2005 more than a thousand papers utilising this SRC technique have been published. Even though the basic principle of all these techniques is the same, there are subtle differences between them. For example the CoMFA and GRID use computed interaction energies between the ligand and a probe, usually an atom or small molecule, as descriptor values. The GRID technique even allows the probe to re-orient itself so that the interaction energy between it and the ligand is optimal. This enables the usage of asymmetrical probes. On the other hand the SOMFA utilizes descriptors which are directly derived from the intrinsic properties of the molecules and no probe is required.

The Comparative Molecular Active Site Analysis (CoMASA)[201] does not have a global grid but instead it relies on a set of pseudoatoms placed at critical points derived from the aligned molecule set via cluster analysis. This reduces the number of variables thus reducing the computational demands of this method but it also eliminates the problems arising from the regular grid, namely the sensitivity to rotation and translation.

Comparative molecular surface analysis (CoMSA)[202-206] computes the descriptor values on the surface of the molecules and then this 3D surface is transformed into a 2D plane. Self-organizing maps (also called Kohonen maps) are used to distil the important information from the descriptor after which a standard PLS analysis is performed. In addition to CoMSA, there is a highly similar technique proposed by Hasegawa et al[207-209] which more accurately preserves the information about the spatial relationships between important molecular features.

One interesting and very diverse class of 3D descriptors are the similarity index based techniques such as Comparative Molecular Similarity Analysis (CoMSIA)[210-213] which use the similarity indices originally formulated for alignment of molecules as descriptors in a SRC analysis. Among others the molecular similarity indices proposed by Carbo[214] and Hodgkin[215] have been also applied to QSAR[216-219]. The actual properties of this class of SRC techniques are dictated by the formulation of the index used, but in general one can say that the indices are usually not very sensitive to small changes in structure and therefore they may not be optimal descriptors.

Also many different descriptors based on quantum chemistry, such as TQSI[220], MQSM[221-228] and QS-SM[229-231], have been developed, most of which are based on the molecular orbital (MO) approach[232] while others favour the density functional theory[233-240]. Many applications of the quantum chemical descriptors have been reported in the literature but all in all it would seem that these descriptors are not clearly superior when compared to ones derived from molecular mechanistic models[230,241-260]. However there are some quantum chemical descriptors such as polarisability and hardness which have proven to be very useful in SRC studies[237,261-263].

All the 3D techniques thus far presented require molecular alignment. However, a large number of descriptors completely circumvent the problems of alignment by reducing the 3D structure into a rotationally and translationally invariant form[20,102,107,220,232,264-273]. In other words these descriptors are based on the 3D structure, meaning the conformation and configuration of the molecule, but are not sensitive to its position or orientation in 3D space. These descriptors are, quite logically, called alignment-free descriptors, but they could also be called "2½-dimensional" as they are sensitive to the 3D structure of the molecule, but they do not directly depend on global 3D space[20,198]. On the other hand, the descriptors and the results of the models built using these descriptors are often more difficult to interpret than is the case with true 3D descriptors[20].

GRIND[20] is an alignment-free SRC technique which takes a set of 3D molecular descriptor grids computed with GRID[15], CoMFA[14] or any other similar technique and first smoothes them and subsequently transforms them using an autocorrelation functions to form the alignment-free descriptor. The GRIND descriptor has been applied to several SRC datasets with varying results[274-276]. Recently an extension of the original technique aiming to improve stereospecificity

of the descriptor by inclusion of the 3D structure motifs was proposed[197]. Unlike many other alignment-free techniques, GRIND enables the user to correlate the alignment-free descriptor with the original grid and thus enables limited visualisation of the results. Like GRIND, the alignment-free techniques proposed by Broto[264], Gasteiger[265] and Clementi[277] use autocorrelation functions to create co-ordinate independent SRC descriptors. They have certain similarities with GRIND and each other, but they use different source descriptors and transformation algorithms and should therefore be considered independent methodologies.

COMPASS[278,279] is an inventive technique that uses iterative optimisation and neural networks to discover the optimal alignment, called a "pose", along with a more traditional SRC model. The first step when building a COMPASS model is the generation of a set of molecular conformations aligned using a template conformation. This ensemble is then covered with descriptor points and few points are also placed outside the ensemble surface. A COMPASS descriptor, which also includes hydrogen bond donor and acceptor variables, is evaluated using the set of points generated earlier and a model correlating the pose and biological response is generated using a neural network. This model is then used to generate better poses for the molecules and the whole process is repeated until the optimisation converges. Finally a COMPASS model correlating the descriptor with the activity is generated which can then be used to align an unknown molecule and subsequently predict its activity. Thus, the COMPASS methodology uses an internal co-ordinate system and combines superposition and descriptor evaluation steps. As such, it could be considered to belong to the alignment-free techniques, even though the actual descriptor is based on the aligned set of molecules, as it does not require pre-aligned molecule set.

The Weighted Holistic Invariant Molecular (WHIM) descriptors use principal component analysis (PCA, see page 39) to transform the centred molecular co-ordinates into a new system defined by the three primary molecular axes. The resulting descriptor is invariant in regard to both translation and rotation due to the centring and the uniqueness of PCA solution, and thus the molecular alignment is not required. Several different weighting schemes utilising the properties of atoms, such as mass, van der Waals volume and many others, have been proposed[280-286].

Comparative Molecular Moment Analysis (CoMMA)[267,268] interprets the value of an atomic property as "mass" and uses it to find the centre of "mass" for a molecule. Then the atomic property and the distance from the centre of mass can be used to compute the "moment" associated with the property. From the individual moments it is possible to find the principle moments which form a 3D co-ordinate system which depend only on the structure and properties of the molecule and thus the CoMMA does not require molecular alignment as the moment axes can be used to implicitly correlate the molecules. The original CoMMA descriptor is formed from 14 different variables (eq. 16) which include the molecular weight (MW), the principle moments of inertia (I_1-I_3), total dipole and quadrupole moments (μ and Q), the components of dipole moment (μ_1- μ_3), the displacement vector between the magic point of the molecule (i.e., centre of dipole) and centre of mass in the inertial co-ordinate system (for details see Silverman and Platt[267]) and finally the two components of the quadrupole moment. And in addition to these steric and electrostatic descriptors the CoMMA principle can be used to compute an

alignment-free descriptor of any atomic property and an extension providing moments of lipophilicity has gained some popularity[268,287].

$$CoMMA = \{MW, I_1, I_2, I_3, \mu, Q, \mu_1, \mu_2, \mu_3, d_1, d_2, d_3, Q_{11}, Q_{22}\} \quad (16)$$

The spectral descriptors are a group of intrinsically alignment-free 3D descriptors which have gained some popularity within the SRC community. They differ from the spectrum-based techniques introduced in the 0D sections as they use computational techniques to generate the data containing the peak positions and intensities. These values are transformed into a spectrum using a Gaussian smoothing kernel. The actual descriptor is derived by computing the kernel sum at predetermined points along the pseudospectrum using equation 17 where the k is the position of descriptor, the λ_0 is the position of the smoothing kernel and the σ is the standard deviation, or half-height width of the kernel. The spectral techniques include EVA[269,270,288,289] which is based on the IR frequencies and the Electronic EigenVAlue (EEVA)[272,273] which uses the energies of the semi-empirical molecular orbitals to generate the pseudospectrum.

$$EVA_k = \sum \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-\lambda_0)}{2\sigma^2}} \quad (17)$$

2.2.3 Beyond 3D

A view of a rigid molecule fitting to a rigid receptor, which underlies the 3D SRC analysis formalism, is an extreme simplification of the real dynamic nature of the molecular recognition. As the successful SRC models indicate, there are many cases in which this approximation can be done without major delerious effects. Nevertheless, there are some cases where the dynamic effects are considerable[95,290,291]. Also in 3D SRC one must decide the active conformation, and in many cases also the orientation, of the molecule in the binding pocket. For rigid molecules this is usually not a problem, but for more flexible molecules it can be very difficult to decide the active conformation. Also for very diverse sets, especially if the receptor structure is not known, it is extremely difficult to select a binding orientation. The 4D-QSAR paradigm circumvents the problems caused by the selection of active conformation and orientation by representing the ligand molecule with an ensemble of conformations thus mimicking the dynamic nature of the binding phenomena[291].

One of the earliest positions for an ensemble descriptor is the so-called cell occupancy method where the SRC grid is composed from small cells and a collection of ligand conformations are placed into this grid. Then the occupation densities, meaning simply how many ligand atoms intersect the cell, are evaluated. Instead of just measuring the simple occupation this method can also be used to formulate more specific descriptors, such as hydrogen bond donor and acceptor fields. After the original paper by Hopfinger et al[290], several applications of the 4D-QSAR paradigm have been reported in the literature[292-312]. It has become clear that in certain cases the ensemble methods have clear advantage over the traditional 3D methods. Unfortunately it is also clear that this new approach also generates a new set of problems and more work is required before the 4D methodology becomes widely adopted[7,312,313]. Furthermore, two different

types of 4D SRC methodologies exist, one which closely resembles the 3D SRC methods as it does not require a known receptor[304,306-308], while another formulation utilises the additional information inherent in the receptor structure[305,308,312,314,315].

If one includes the receptor structure in the SRC analysis it means that one gets additional information which can be used to improve the accuracy of the model but at the same time one gets an additional degree of freedom, *viz*. the conformational flexibility of the receptor, which should also be taken into account. Therefore one should extend the original 4D formalism to include several receptor conformations thus creating a 5D approach[95,291]. Some successful applications of this new SRC methodology have been reported, but it is clear that it will take some time before these techniques becomes widely accepted[314,316,317]. While this 5D methodology is still in its infancy the team led by A. Vedani, who originally proposed the 5D SRC and wrote an implementation in the program QUASAR, have further increased the degrees of freedom to create a 6D SRC[318]. Only time will tell whether this new techniques yields sufficient increase in predictive power to offset the rapidly increasing computational cost.

2.3 Descriptor post-processing

After the descriptor values have been calculated for all of the molecules they are usually collected into a matrix (X) where each row corresponds to a sample and each column is a different descriptor variable. A similar matrix (Y) can be composed from the observed values. However, as the vast majority of SRC analyses have univariate observed values Y is most often a vector rather than a matrix. Even if the SRC analysis has a multivariate observed values there is controversy whether it is better to have a single multivariate analysis or a set of univariate analyses[319,320].

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{18}$$

where **n** is the number of samples, in this case compounds, and **m** is the number of descriptor variables.

Due to the nature of SRC analysis, it is common that there are more descriptor variables than there are samples. In the case of grid-based 3D descriptors such as CoMFA there can be literally thousands of variables while the number of samples is usually in the range of 20-50. To compound the problem many descriptors have considerable internal correlation. In other words many descriptor variables are more or less co-linear, which in turn causes problems for many statistical analysis methods. One should also note that a for set of highly co-linear variables, as opposed to a single variable, there is only a modest contribution to the information content of the descriptor, but as each of the variables contains noise, the co-linearity effectively reduces the signal to noise ratio of the descriptor. This problem is also referred as internal correlation, meaning that the variables within or between descriptors correlate. Therefore it makes little sense to simply create a super-descriptor by gluing together all available descriptors as it would not significantly increase the predictive power of the model[319-321].

Naturally, the more complex statistical tools, which use weight factors, can be used to discard poor variables but they are usually capable of discarding only the most spurious variables and therefore several methods have been devised to identify and discard the least informative variables from the descriptors. These methods can be divided into two major categories, namely *variable reduction* and *variable selection* approaches, based on the data used to guide the process. In the variable reduction the X matrix is analysed and variables with high internal correlation or very weak variance are deleted thus reducing the number of variables while preserving the maximum amount of information. On the other hand, variable selection methods also take in account the observed values (Y) when selecting the best variables. The region focusing often used in conjunction with CoMFA, is a part of the variable selection methods as it uses the R^2 value of the final model to guide the selection of certain CoMFA descriptor variables, which are given a greater weight when building the final statistical model.

In CoMFA the descriptor variables are derived using a 3D grid and thus the variables have a unambiguous location in 3D space. Thus the sets descriptor variables correspond to a region, or regions, of space which explains the name of the technique as it quite literally focuses on regions of interest.

Regardless of whether or not the *X* matrix has been treated with some variable reduction or variable selection methodology, the resulting matrix is usually subjected to two statistical pre-treatments, namely *mean centring* and *variance scaling*. In mean centring for each of the descriptor variables its average over all samples (column of matrix *X*) is evaluated and subsequently subtracted from the same variables so that the mean of the variables (column) becomes zero. This data pre-treatment enables certain mathematical shortcuts and is required by some statistical algorithms. As the mean centring only affects the absolute values of the variable and leaves the relative positions unchanged it usually does not have any negative impact on the efficacy of the statistical analysis. Variance scaling is used to normalise each of the descriptor variables to unit variance which ensures that all variables have equal weight in the statistical analysis. However, in some cases the differences in the ranges of variables can act as intrinsic weight factors and the variance scaling, which removes them and actually reduces the accuracy of the statistical model.

2.4 Model building and statistical analysis

In this section many of the statistical techniques used in quantitative SRC analyses are introduced. The primary emphasis of the discussion is on the linear regression techniques, as they still are the method of choice for quantitative SRC analysis. As was the case with the DYLOMMS, the evolution of linear regression has in many cases dictated the success of new SRC paradigms and therefore the choice of statistical method is crucial part of the analysis. This overview of statistical analysis tools follows the chronological progression of linear regression starting with the relatively simple Multiple Linear Regression (MLR) and progressing through the principal component based methods, such as Principal Component Analysis (PCA) and Partial Least-Squares (PLS), to non-linear and non-parametric regression including artificial neural networks.

In recent years it has been demonstrated that it is not necessarily best to use the most powerful statistical methods available. The main critique against those powerful techniques is that while they undoubtedly are capable of detecting weaker correlations and to generate a more flexible model, the increase in true predictive power is often rather modest. The reason for this is that the powerful techniques learn the set of molecules used to generate the model, the so-called *training set*, too well and lose their ability to accurately predict new compounds[103,322-324]. Due to these problems a set of simpler statistical analysis techniques has been introduced, of which the k-Nearest Neighbours approach has gained greatest popularity[102,104,207,322,325-338] and which could, due to its computational simplicity, high predictive ability and the robustness of its models, one day rival linear regression as the method of choice for structure response correlation analyses[102,322].

One should also note that the classification techniques used in non-quantitative SRC analyses are not discussed in this overview as this work is based on the quantitative structure response analysis methodology and therefore, strictly speaking, the non-quantitative methods do not fall within the scope of this work. For a recent review of classification techniques commonly used in the SRC analyses, the reader is referred to an article by Mazzatorta et al[339].

In this overview the uppercase letter denotes a matrix whereas a lowercase letter represents a vector or a singular value. The apostrophe (') is used to denote a matrix transpose. Also, the observed values are called *dependent variables* and the matrix containing them (Y) is called *dependant block* as is customary in the literature discussing regression methodology. The logic behind these names is that the Y values are assumed to linearly depend on the *independent variables* contained in the X matrix, which is often called *independent block*. How can the descriptor variables in the X matrix be independent as they depend on the structure and the descriptor used to evaluate them? The values in X do indeed depend on the structures and descriptors, but in the context of regression analysis, they do not depend on anything and are therefore called independent variables.

2.4.1 Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) is the oldest and simplest of linear regression methods. It is, however, still quite useful in classical SRC analysis with a small number of highly orthogonal variables. The basic equation of MLR model is shown in equation 19, where B_{MLR} is a matrix of regression coefficients computed with equation 20 and the E is a matrix of residuals (errors). After the B_{MLR} has been computed, an estimate of the Y, denoted by $\hat{Y}$, can be computed for an arbitrary set of independent variables by equation 21. The greatest weakness of the MLR analysis is that if the independent block X contains highly co-linear variables the inverse of $X'X$ may not exist (eq. 20) and the MLR fitting will fail. Also if the number of variables is greater than the number of samples the MLR will not yield a unique solution but rather a set of possible solutions. This limits its usefulness in SRC models as the number of descriptor variables tends to be much higher than the number of samples.

$$Y = XB_{MLR} + E \tag{19}$$

$$B_{MLR} = (X'X)^{-1}X'Y \tag{20}$$

$$\hat{Y} = XB_{MLR} \tag{21}$$

2.4.2 Ridge regression (RR)

Ridge regression is a biased regression technique which can be used instead of the MLR for highly co-linear or undetermined data sets. The difference between ridge regression (eq. 22) and standard MLR (eq. 20) is the inclusion of a bias term kI where the k is a non-negative "ridge" constant and the I identity matrix. If the k=0 the ridge regression is identical to MLR and when k increases it will introduce increasing amount of bias in to the regression.

The biased regression techniques such as ridge regression have received a lukewarm reception at best[340]. This is probably due to the difficulty involved in finding the optimal bias, though there have been some success stories where the biased regression has out-performed other regression techniques[340,341]. Normally k is optimised using a grid search and a large ensemble of ridge regression models in order to achieve sufficient statistical reliability. This means that fitting a RR model requires considerable amount of computing power, though recently some reliable methods for fast estimation of k has been proposed[342,343].

$$\hat{B}_{RR} = (X'X + kI)^{-1}X'Y \tag{22}$$

2.4.3 Principal Component Analysis/Regression (PCA/R)

In *Principal Component Analysis* (PCA) the independent block *X* is decomposed into **a** *principal components* (PCs), described by the *t*- and *p*-vectors (eq. 23) containing the scores and the loadings of samples and variables, respectively. The decomposition of *X* can also be interpreted as a co-ordinate rotation where the original axes, defined by the descriptor variables, are replaced with a new set of orthogonal, and thus not co-linear, axes called principal components (PCs). As the PCs are generated so that each explains the maximum amount of the residual variance not yet explained by preceding PC, one can quite easily see that the relative importance of new PCs decreases as more and more components are extracted. The maximum number of PCs that can be extracted is the same as the number of descriptor variables in the *X* matrix. On the other hand, as the data always contain some inherent error and X-variables correlate, it is usually the case that only relatively few PCs contain all relevant information and the rest can be discarded as they contain increasing amounts of noise. For a short discussion about the procedure to select the optimal number of principal components, please see page 50.

$$X = t_1 p_1' + t_2 p_2' \ldots + t_a p_a' + E = TP' + E \tag{23}$$

The loading vectors have as many elements as there are original variables and they contain the cosines of angles between the principal component and the original variables. For example in the case presented in the Figure 4 the loading vectors are $p_1 = \{\cos\theta_{11}, \cos\theta_{12}\}$ and $p_2 = \{\cos\theta_{21}, \cos\theta_{22}\}$ for PC_1 and PC_2, respectively. As the loadings vector is usually normalised to unit length one can also interpret the loadings vector as a unit vector defining the new axis based as a linear combination of the original variables. The score vectors contain the projections of the each sample onto the principal components and they therefore contain as many elements as there are samples. The scores can also be interpreted as the co-ordinates of the samples expressed in system defined by the principal components instead of the original variables (see t_{16} and t_{23} in Figure 4).

In order to perform a *Principal Component Regression* (PCR) one must derive a matrix of regression coefficients from the results of the PCA. This can be achieved by collecting the loading vectors into a matrix where each column corresponds to an original loading vector (eq. 24). Then, by using the *P* matrix one can transform the original descriptor *X* into the principal component co-ordinate system and create a new descriptor matrix *T* (eq. 25). When the matrix *T* is used instead of the *X* matrix in equation 20 one can easily derive the regression coefficients (eq. 26) as the principle components are, by definition, orthogonal there is no matrix inversion problem.

For a more detailed discussion and list of applications and different implementations of principal component analysis and principal component regression the reader is referred to a book entitled "*A User's Guide to Principal Components"* by J. E. Jackson[344].

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1a} \\ p_{21} & p_{22} & \cdots & p_{2a} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \cdots & x_{ma} \end{bmatrix} \tag{24}$$

$$T = XP \tag{25}$$

$$\hat{B}_{PCR} = (T'T)^{-1}T'Y \tag{26}$$

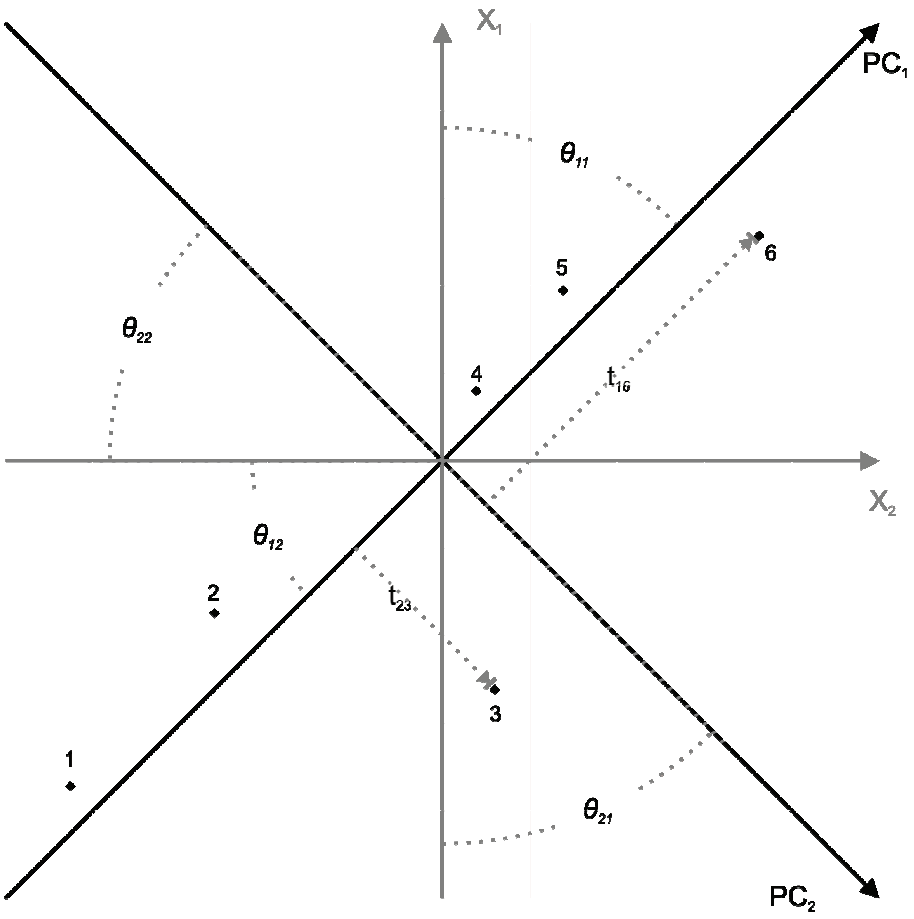


Figure 4. Schematic representation of Principal Component Analysis (PCA). The x_1 and x_2 are the original independent variables and the PC_1 and PC_2 are the two principal components which optimally describe the variance of the data. Cos θ_{11} and cos θ_{12} are the loadings for PC1 and θ_{21} and θ_{22} are the loadings for PC2. T_{16} and t_{23} are the 6th score of PC1 and the 3rd score of PC2, respectively.

2.4.4 Partial Least Squares (PLS)

The *Partial Least Squares* (PLS) is an advanced regression methodology, originally designed for econometric applications, which has also been extensively utilised in chemometric applications[345-347]. In PLS analysis the principal components of both the X and Y blocks are decomposed using PCA (eq. 27) so that the scores of the dependent block $u_1 \ldots u_a$ form a n by a matrix U. In similar manner the scores of the independent block form $t_1 \ldots t_a$ matrix T. The main difference between the PCA and PLS is that the scores of the dependent and independent blocks are mixed and therefore the X- and Y-blocks will provide information about each other. This ensures that the X-scores are at all times strongly correlated with the Y vector which is not necessarily the case with PCA/R. The mixing also improves the tolerance of the PLS for highly colinear independent variables. This mixing has the side-effect that the PLS components are not necessarily orthogonal and therefore it is necessary to introduce a new matrix W (m by a) which contains the weight factors ($w_1 \ldots w_a$) for the independent variables. The matrix W and the loadings are used to derive the regressions coefficients for the final orthogonal PLS components (eq. 28)[319,320,348,349].

$$\begin{aligned} X &= TP' + E \\ Y &= UQ' + F \end{aligned} \tag{27}$$

$$\hat{B}_{PLS} = W(P'W)^{-1}Q' \tag{28}$$

Table 2 presents a detailed pseudocode of a PLS analysis methodology derived from the Non-Iterative Partial Least Squares (NIPALS) PCA algorithm. The NIPALS PLS is by no means the best, or only implementation, but it is easier to follow and more intuitive than the more efficient implementation and therefore it is used here as an introduction to the principles of PLS regression.

To begin the PLS analysis let $F_0 = Y$ and $E_0 = X$ and then the first principal component is extracted using the algorithm defined in Table 2. First u is initialised (line 1) and then w is computed and normalized (lines 2-3). Then w is used to compute t (line 4) which in turn is used in computation of normalised q (lines 5-6). It should be emphasised that in the computation of the t and q vectors the weight vector w contains information about the dependent block, thus mixing the independent and dependent blocks. Finally a new u is derived and a check whether the change in t is smaller than an arbitrary convergence criteria ε is made (lines 7-8). If the PLS is computed for a case with univariate dependent block, the lines 5-8 can be substituted with q=1 and the iteration is unnecessary. To obtain more principal components one must evaluate new E and F matrices using equation 29 and then re-iterate the PLS algorithm using these new matrices.

$$\begin{aligned} E_{n+1} &= E_n - t_n p_n' \\ F_{n+1} &= F_n - u_n q_n' \end{aligned} \tag{29}$$

As mentioned earlier, there are many efficient implementations of the PLS methodology, such as SIMPLS[350], SAMPLS[351], SVDPLS[352] and others[319,321,348,353-358]. Though some of these methods yield slightly different results, they are variations on a theme, with slightly different computational complexity and limitations. For example some PLS implementations are restricted to univariate *Y*. For a more detailed description on the mathematics behind the PLS, the reader is referred to articles by Wold[346], Höskuldsson[348], and Geladi[320].

Table 2. Detailed pseudocode for the evaluation of j[th] principal component of a PLS implementation based on the NIPALS PCA algorithm. At the end of iteration the local vectors u, w, t and q become the global vectors u_j, w_j, t_j and q_j which are also placed in the U, W, T and Q matrices.

1 $u = F_j$

2 $w' = \frac{u'E_j}{u'u}$

3 $w = \frac{w}{\|w\|}$

4 $t = \frac{E_j w}{w'w}$

5 $q' = \frac{t'F_j}{t't}$

6 $q = \frac{q}{\|q\|}$

7 $u = \frac{F_j q}{q'q}$

8 if $\frac{\|t - t_{prev}\|}{\|t\|} > \varepsilon$ goto 2

2.4.5 Non-linear and non-parametric regression

To clarify the concepts of non-linear and non-parametric regression one should consider a case where the observed values (y) are dependent only on a single descriptor variable (x). In this two dimensional case the regression model can be interpreted as a curve which should intersect with all sample points. Similar geometric interpretations can be made for higher dimensionality cases where the regression model forms a (hyper-) surface in the combined dependent/independent variable space. For clarity of presentation, only the highly intuitive 2D cases are discussed. In the case of linear regression the model curve is a line (eq. 30). At first glance, it would seem that linear regression generates an overly simplistic model and a more complex curve would considerably increase the predictive power of the model. Instead of the line one could use a parabola (eq. 31) or higher order polynomial (eq. 32) yielded by quadratic and polynomial re-

gression, respectively[359]. Unfortunately the increase in the predictive power is rather modest and as the regression model becomes increasingly complex its sensitivity for any, and all, noise in the data also increases. This is due to the fact that as one increases the degrees of freedom in the regression model without increasing the available information, in this case meaning the number samples, the less well defined the fitting becomes. In other words, the more information one tries to extract from a limited set of samples the more one has to infer from the limited sample and thus the reliability of the model is reduced. Regardless of these problems several success stories have been reported[360-367] and there is some evidence which suggests that non-linear regression could in most cases be superior to linear regression[368,369]. However, at the moment, the linear regression is still the method of choice when performing SRC analyses. For more details the reader is referred to publications by Wold et al[369] and also to Tang and Li[368].

$$y = ax + b \tag{30}$$

$$y = a_1 x^2 + a_2 x^1 + b \tag{31}$$

$$y = a_1 x^k + a_2 x^{k-1} \ldots + a_{k-1} x^2 + a_k x + b \tag{32}$$

Unlike the linear and non-linear regression methods, non-parametric regression does not try to form a single parametric curve, rather it generates a surface defined by the known points, meaning the samples, and suitable weight function. The actual derivation of the non-parametric regression model is quite complicated and the mathematical details do not fall into the scope of this work. A concise primer into the mathematics of the non-parametric regression, along with a host of references is presented in paper by Constans et al[370].

Non-parametric regression bears a resemblance to the k-nearest neighbours method (see page 46) as the predicted activities are computed using the activities of neighbouring molecules. The use of a kernel functions gives more flexibility to the definition of a neighbour and thus also to the non-parametric regression[370,371]. Due to its flexibility and robustness, non-parametric regression is becoming a useful tool for QSAR as it does not require explicit definition of regression model and can model diverse systems[372]. Unfortunately, non-parametric regression is computationally demanding and therefore requires more resources than standard linear regression tools. However, increasing power of the computer hardware is rapidly diminishing the impact of this difference and thus making non-parametric approaches more accessible[370,372]. For more information about non-parametric regression algorithms the reader is referred to articles by Hirst et al[371], Constans et al[370], McNeahy et al[372].

2.4.6 Artificial Neural Networks (ANN)

As the name suggest the artificial neural networks are mathematical tools which try to mimic the biological neurons and the networks they form. This non-linear form of statistical analysis has stirred up great interest in the SRC community but as always many problems have also been reported[325,364,373-397]. The basic concepts of the ANN approach are quite intuitive as there are clear biological counterparts for the mathematical constructs.

The artificial neuron (Figure 5A) consists of an arbitrary number of inputs (I) which corresponds to the dendrites of a biological neuron. These are scaled by weight factors (w) and fed to a function *f*, called the activation function, which is used to evaluate the output (O) acting as the "axon" of the artificial neuron. An artificial neuron could, in principle, have a number of independent outputs. However, this is usually not the case as neural networks would most likely become unmanageable due to the topological complexity[398]. Even though the neuron has an activation function and a set of weights at the inputs, the capabilities of single neuron are still rather modest. Therefore one could say that even thought the neuron is stupid, the network of neurons is collectively smart. The ANN could be interpreted as a form of cluster computing where a massively parallel system of simple unit performs complex operations. As the "smartness" of the system depends on the topology of the network, meaning the number of neurons and the way in which they are organised and interconnected, the initial layout of the networks is crucial as it ultimately defines capabilities of an ANN[398]. There are very few theoretical limitations to the structure of an artificial neuron or to the complexity of a neural network but, in practice, as one must use a sequential computer to mimic the multiply parallel network, there are only a handful of truly viable topologies. The most popular topology is the *feed-forward* network where the neurons are arranged in tiers which are evaluated sequentially so that the neurons pass their outputs as input for the next tier of neurons. There are also many ways in which the tiers can be interconnected but usually each neuron is connected to all neurons of the previous tier thus forming the so-called *fully connected* network. A schematic representation of a simple ANN formed using a fully connected feed-forward topology is presented in the Figure 5B. The input data, or in the case of SRC analysis the descriptor, is fed to the input stage (I_1-I_5) which is connected to a first tier of processing neurons (N_{11}-N_{13}). A tier of neurons which do not belong to the input or output stage is often referred as a *hidden layer*. In the Figure 5B there are two hidden layers of neurons (N_{11}…N_{13} and N_{21}…N_{22}) along with the input (I_1…I_5) and output stages (N_{31}).

A statistical analysis using an artificial neural network can be divided into three separate phases. First one must set up the ANN by deciding the topology of the artificial neural network and the activation function used in neurons. One should also initialise the parameters in the neurons to random values at this phase. In the second phase the neural net learns by example just as a human does. The data are fed to the networks and the parameters are adjusted using a suitable learning function. The learning can be unsupervised where the network strives to uncover internal structure from the data. In the supervised learning the output of the networks is compared to the known values associated with the data and the difference, or error, is used to train the networks[398].

The training of ANNs is a very delicate process and it is often necessary to take particular care to find the optimal training regimen. Thus it is often necessary to perform extensive validation in order to ensure that the network is indeed optimal also for a more general case than just training[393].

Traditionally the ANNs are considered to be black boxes because the information about the relative importance of the input variables is present in a holographic fashion in all of the weight factors. It is very difficult to separate a neurons input weight factor into different factors corresponding to the variables and thus it is also nearly impossible to gain insight into the underlying reasons for the observed regularities. Nevertheless, some recent papers indicate that it might be possible to extract information from ANNs[399,400]. For a detailed introduction to the use of artificial neural networks in chemistry the reader is referred to the *Neural Networks in Chemistry and Drug Design* by Zupan and Gasteiger[398]. Also see papers by Sutherland et al[7] and Agfafiotis et al[395].

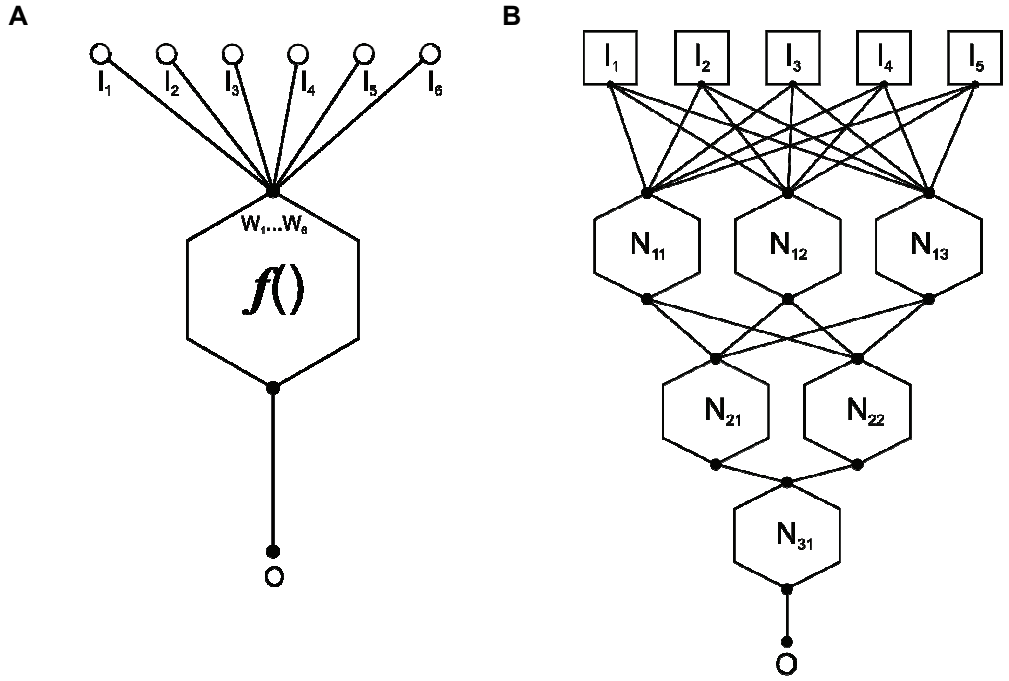


Figure 5. Schematic representation of an artificial neuron (A) and a neural network (B).

2.4.7 k-Nearest Neighbours (kNN)

K-Nearest-Neighbours (kNN) analysis is probably the simplest and computationally easiest classification/regression technique in existence[401] as it relies on the simple assumption that the value of an unknown sample can be predicted using the values of its nearest neighbours. kNN does not make any other assumptions about the nature of the relationship between the value of a point and its descriptor. This makes it highly suitable for unevenly distributed datasets as each cluster of samples is automatically used to predict similar samples and the remote clusters do not interfere as is the case with many other techniques[329,402].

In its most common form, this technique needs only one parameter: k, which indicates the number of neighbours used (usually k=3,5,7,9…). It also does not require a specific teaching step and this lack of a separate model building step enables the easy expansion of training set without any modification to the model. Additionally, due to the lack of model building, the kNN technique usually generates a model which may have slightly weaker predictivity than other methods, but on the other hand, they usually have more robust external predictivity than the regression models[104,394]. For classification purposes a majority voting among the neighbours is the most popular approach whereas in quantitative kNN, the distance weighted average of the neighbour values is used. Due to its simplicity and robustness the use of kNN has gained momentum among the practitioners of the SRC analysis and to a certain extent it has surplanted traditional regression methods, such as PCR and PLS[101,102,394].

Even though the actual mechanism of the kNN is very simple, it is often very difficult to gain insight to the reasons for the observed regularities. In this sense the kNN functions as a black box and can not be used to gain additional information about the physico-chemical causes behind the observed differences. On the other hand, one can quite easily create a distance matrix which can be used to cluster the data points and this information can in turn be used in conjunction to the original data to try to elucidate a model[403]. Even though it is possible to use an arbitrary distance-metric in kNN analysis to identify the nearest neighbours, the simple multidimensional Euclidian distance is usually selected because of its computational simplicity. Due to the nature of the kNN analysis it requires the data to be variance scaled and mean centred, or the different value ranges may act as implicit weighting factors, thus unduly biasing the results. One should also note that as the basic kNN assigns equal weight to each descriptor variable it is very susceptible noise and co-linearity[404]. Therefore it is usually prudent to use some form of variable selection prior to the kNN analysis[405]. Alternatively, a complex distance-metrics or a set of weight factors for variables and samples can be applied in order to improve the tolerance for co-linearity and noise but at the same time one usually loses the simplicity and robustness of the kNN analysis.

2.5 Model validation

After the statistical analysis of a SRC model is complete, it is essential that one can derive some quantitative measure of the predictive power and goodness of the fit of the new model. For the SRC analyses which only seek to classify the samples there is a single score, the *accuracy* of the classification (eq. 33) ranging from 0.0 to 1.0, which contains all the necessary information to assess the predictive power of the model.

$$accuracy = \frac{correct\ classifications}{n} \tag{33}$$

For quantitative models one needs more complex scores in order to estimate fully the power of the model. These scores include: the predictive residual sum of squares (*PRESS*, eq. 34, range: 0.0–∞), sum of squared deviations (*SSD*, eq. 35 range: 0.0–∞), standard error (*SE*, eq. 36, range: 0.0–∞) and squared correlation coefficient (R^2, eq. 37, range: -1.0–1.0)

$$PRESS = \sum \left(y_{obs} - y_{pred} \right)^2 \tag{34}$$

$$SSD = \sum \left(y_{obs} - \overline{y}_{obs} \right)^2 \tag{35}$$

$$SE = \sqrt{\frac{SSD}{n}} \tag{36}$$

$$R^2 = \frac{\left(\sum \left(y_{obs} - \overline{y}_{obs} \right) \left(y_{pred} - \overline{y}_{pred} \right) \right)^2}{\sum \left(y_{obs} - \overline{y}_{obs} \right)^2 \sum \left(y_{pred} - \overline{y}_{pred} \right)^2} \tag{37}$$

When estimating the predictive ability of a SRC model it is necessary to distinguish two classes of predictive power, namely the *internal* and *external predictivity*. The internal predictivity measures how accurately the model can predict the set of compounds which was used to build the statistical model, in other words the *training set*, while the external predictivity tries to measure the model's predictive power for molecules which it has never seen before. Of the two, the external predictivity is more important as it more accurately indicates the true power of the model in a realistic situation where it is used to predict unknown compounds[401].

It is a well known fact that even though a SRC model can be used to predict any molecule the descriptor evaluation can handle, the reliability of prediction can only be guaranteed for those molecules which resemble the compounds used in the training set. This is due to the fact that the cluster of samples forming the training set covers only a fraction of the whole descriptor space. Therefore the correlation between the structure and response is also well defined only within this descriptor subspace where the model can function in an interpolative fashion. The errors involved in this estimation are usually small because interpolation is in essence a form of

inductive reasoning. On the other hand when the predicted compounds begin to differ from the training set compounds, it will inevitably leave the area of well-defined structural space. When this happens, the model must begin to work in an extrapolative mode and to use a deductive reasoning to generalise the correlation observed for a small set of molecules to all molecules[406]. Therefore, the true external predictivity of a model can never be reliably approximated for an arbitrary molecule. However, for a limited set of compounds, which resemble the training set, the external predictivity can be estimated. One way to estimate the external predictivity of the new SRC model is to use a set of samples called *external test set* or simply *test set* which contains samples that have not been used to train the model[17].

Unfortunately, the limitations in estimating the true external predictivity are not the only problems one must take into account while estimating the true power of a SRC model. There is a far subtler problem involving the way in which the model is trained. For any statistical method which uses the training data to infer a model it is usually possible to generate a carefully fitted model which can predict the training set with almost perfect accuracy. Unfortunately, at the same time one usually loses a great deal of external predictivity because the model which can predict certain set of compounds with perfect accuracy is too specialized to accurately predict compounds that differ from the set which was used to optimise the model. This kind of over specialised model is often referred as an *over-fitted* model because it depends too much on the specifics and peculiarities of the training set. As there are over-fitted models, there are also *under-fitted* models which do not fully utilise the information available in the training set and therefore do not gain optimal performance. Thus one must always balance the trade-off between the efficient use of the training set and the generality of the model[324,394,407,408].

For external test sets the following scores are usually evaluated: squared correlation coefficient (R^2_{ex}, eq. 37, range -1.0–1.0), predictive residual sum of squares for the external test set ($PRESS_{ex}$, eq. 34, range: 0.0–∞), mean absolute deviations ($|\Delta|_{av}$, eq. 38, range 0.0–∞), the sum of squared deviations between the values of samples in the test set and the mean activity of the training set samples (SSD_{ex} , eq. 35, range 0.0–∞), and standard error of prediction (*SDEP*, eq. 39, range: 0.0–∞)

$$|\Delta|_{av} = \frac{\left|\sum \left(y_{obs} - y_{pred} \right)\right|}{n} \tag{38}$$

$$SDEP = \sqrt{\frac{PRESS_{ex}}{n}} \tag{39}$$

Perhaps the most powerful, and also most under-used, score for the assessment of the true predictive power of a SRC model is the predictive R^2-score ($Pr\text{-}R^2$, eq. 40, range -∞–1.0), which indicates predictive power of the fitted model as compared to the naïve model where predicted value of every sample is equal to the mean of the values of the training set. Negative values indicate that the fitted model is inferior to the naïve model and should therefore be discarded whereas positive values indicate that the fitted model has some predictive power.

$$Pr - R^2 = \frac{SSD_{ex} - PRESS_{ex}}{SSD_{ex}} \tag{40}$$

However using a single external test set is, for a statistical point of view, not very reliable methods of testing the external predictivity of SRC model as the selection of compounds into the training and test set can lead to a considerable bias in the results. Therefore, instead of using a single external set one can also exclude a part of the original training set and use the rest to generate a model to be used for the prediction of the excluded samples. This procedure is called *cross-validation* (CV). It is also possible to perform a CV by using the original training set and still use an external test set. The benefits of the cross-validation procedure are that it is possible to generate a considerable number of different combinations of reduced training sets and corresponding internal test sets leading to considerably better statistical reliability than could be achieved by single test set. The CV also allows one to estimate the statistical robustness of the model as an optimal model would not lose any predictive ability. For all practical datasets the performance of a cross-validated model is always significantly less than the non-cross-validated scores would indicate.

The most common form of cross-validation is *Leave-One-Out* (LOO) where each sample is excluded once from the training set and used as an internal test sets as described earlier. As this process is repeated for all samples the results obtained from the excluded values can be used to estimate the external predictivity of the model. Unfortunately many datasets have a considerable structural redundancy, meaning that even if a certain sample is excluded from the training set a nearly identical molecule can be found, which can severely compromise the reliability of the LOO CV. In part this can be countered by using *Leave-Many-Out* (LMO) or synonymous *Leave-Some-Out* (LSO) cross-validation techniques in which the training set is divided into a larger blocks each containing 5-30% of samples which are in turn excluded just as in the LOO CV. The advantage of this method is that by using a larger block the structural redundancy is not as great a problem as in the case of single molecule. Usually when the LMO blocks are generated it is prudent to choose as balanced set of samples as possible. It makes little sense to exclude a whole family of similar samples and then use the other samples to predict those samples as this means the SRC analysis would have to operate extrapolatively.

For the cross-validated model the following scores, which are analogous to the scores obtained for the non-cross-validated model but use the CV external predictions as y_{pred}, can be evaluated: cross-validated standard error of prediction (S_{PRESS}, eq. 41, range: 0.0–∞) and cross-validated squared correlation coefficient (Q^2, eq. 42, range: -∞–1.0).

$$S_{PRESS} = \sqrt{\frac{PRESS_{CV}}{n - NPC - 1}} \tag{41}$$

$$Q^2 = 1 - \frac{PRESS_{CV}}{SSD_{CV}} \tag{42}$$

where the **n** is a number of samples, the **NPC** is the number of the principal components extracted, or NPC=1 if the analysis technique is not based on principal components. Please note that the SPRESS value is weighted so that it penalises more complex models with high number of principal components, thus reducing the risk of over-fitting.

In addition to their role in determining the predictive ability of a SRC model the cross-validated scores Q^2, and occasionally S_{PRESS}, are also used to decide the optimum number of principal components for PCR and PLS regression models. There is a rule of thumb that the number of principal components (NPC) should not exceed one quarter of the number of samples used to train the regression model, but to objectively decide the exact optimum number of PCs one needs to use cross-validated scores. In the SRC community there has been a lively debate whether one should use the minimum S_{PRESS} or the maximum Q^2 score to guide the selection of optimum number of PCs, but no clear consensus was found and both scores are still in use[401]. For most cases the scores indicate the same optimum number of components but on an occasion the S_{PRESS} favours a slightly smaller number of PCs.

An alternative to the LMO CV is the technique called *bootstrapping* where 20-50% of the samples in the training set are randomly allocated into external test set. The rest of the samples are used to generate a model which is used to predict the values of the aforementioned external test set. It is customary to use LOO CV while generating the bootstrapping models. After the model has been built, the scores describing the internal and external goodness of the fit are computed. When this procedure is repeated thousands of times the sheer mass of the different division will cancel out the exceptionally good and also the exceptionally bad models. Thus the minimum, maximum and average scores found in the ensemble of bootstrapping runs indicate the minimal, maximal and most likely performance of the model.

Unfortunately the problem of estimating the true external predictivity is not the only problem affecting the SRC models. As early as 1972 Topliss warned that when QSAR descriptors with large number of variables are combined with powerful statistical tools, there is a considerable risk of chance correlations[409]. In classical QSAR the descriptor typically contained less than 10 variables and as the training set used often contained 20-50 molecules it meant that the risk of chance correlation was neglible. However, with the advent of more complex topological and electron state descriptors chance correlation became a real pitfall in QSAR analysis. As the grid-based 3D descriptors, with thousands of variables, were introduced the problem of meaningless correlations became acute. As the more powerful and flexible statistical tools are used to generate SRC models, weaker and weaker correlations can be found. This also means that the "signal-to-noise" ratio of the observed correlation is reduced and at some point it becomes impossible to differentiate between a "true" correlation and a spurious correlation created by the noise in the data. To make matters worse, the R^2 or Q^2 guided variable selection techniques further exacerbate this problem, as they endeavour to enrich all correlations and therefore also tend to favour models with spurious correlation. The other possible source for chance correlation is also the enormous number of available descriptors. If many different descriptors are tested for a certain set of molecules, each new set of descriptors increases the chance of finding a spurious correlation between the observed values and the descriptor variables.

Due to the possibility of chance correlations a high Q^2 value is only a necessary, but not sufficient, condition for high predictive ability. Therefore, as Tropsha et al have pointed out[410], an additional test, called Y-scrambling, is required to fully validate the QSAR model and to assess its statistical robustness. In this test, as the name suggests, the observed values of the training set are randomised and a new model is derived using these meaningless data. Also, one should note that it is necessary to perform the variable selection and region focusing separately for each of the scrambled sets as those techniques heavily depend on the both *X* and *Y* matrices.

If the correlation between the descriptor and the observed values is not spurious, there is dependence between the descriptor and observed values, which is severed when the *Y* block is randomised, and thus the correlation will be lost. If, on the other hand, the correlation is spurious it is very likely that a new random correlation can be found for the scrambled data and no significant loss of correlation will be observed. Therefore, if the scrambling of the observed values leads to a full loss of predictive power, one can deduce that the observed correlation was not random. For a full Y-scrambling analysis one should generate thousands of randomised models in order to gain enough mass into the ensemble so that the occasional spikes in predictive ability, caused by poorly scrambled sets, will be averaged out.

2.6 Visualisation and the inverse problem of QSAR

One clear usage for SRC models is the *in silico* prediction of the properties of unknown compounds, but in many cases one would clearly benefit from a reverse SRC model enabling one to predict a structure from activity[1]. In particular there are many molecular discovery approaches, such as rational drug design, which are actually not that interested in the exact responses but strive to find molecular scaffolds and modifications which could be used to synthesise new active molecules. Unfortunately this problem of finding a structure based on the response has proven to be a tough nut to crack and it is often referred as the "inverse problem of SRC"[62,411]. Often the inverse model would be more valuable than the normal SRC one as it could be used to speed up medicinal chemistry and lead discovery. Therefore it is hardly surprising that there has been a considerable interest in the field of inverse SRC but unfortunately the results have remained rather modest.

So the title of this section is actually a bit of a misnomer as it should be *structure elucidation in SRC*. Nevertheless, the techniques which could be used to generate structures that correspond to high or low response remain elusive and therefore one must use the carefully built and validated SRC models to find active and inactive structures indirectly. Naturally one can try to perform a manual reverse engineering by ordering the training and test set molecules by their observed or predicted activity and try to discern any regularities. To take this manual analysis one step further one can then make modifications to the molecules and feed them into the model and see what kind of change in activity the modification caused. All in all this is a rather slow and inefficient procedure. Unfortunately, for many SRC models, it is the only available form of "reverse engineering"[5,412].

On the other hand the grid-based 3D SRC techniques offer an easy and intuitive way to visualise the observed correlations. As each point in the global grid directly corresponds to a single variable it is easy to see that one can use the weight obtained via statistical analysis, in most cases meaning the regression coefficients, to explicitly indicate the relative importance of this point[14,200]. If the 3D SRC uses several fields, such as steric and electrostatic interaction, each of them can be processed separately thus yielding *interaction contour plots* or more simply *contour plots*. Such plots very clearly indicate the areas where more steric bulk increases or decreases activity or where charges have a positive or negative effect on the activity[14]. Even though these plots offer a greatly improved way of interpreting the results of the SRC analysis they are by no means a panacea. The indicated areas of positive or negative influence are often fragmented and therefore many techniques have been devised to refine the raw plots so that the essential information remains and spurious fragments are removed. These smoothed plots are generally less ambiguous and therefore easier to interpret[413-418]. An example of a CoMFA contour plots as generated by Sybyl[94] is presented in Figure 6. The visualisation used is similar to that used in the original paper[14].

This easy and intuitive visualisation is a clear strength of grid-based SRC techniques. This is because as for most SRC techniques it is impossible to create a direct link with structural features and observed activity. Therefore those techniques are limited to the reverse engineering approach described earlier. The reasons for the drastic differences in reverse prediction ability

are caused by the differences in the complexity and redundancy of the descriptors. The evaluation of the grid-based descriptors is a very simple and straightforward process whereas, for example, the value of single EVA spectral descriptor variable depends on a sum of several smoothing kernels which in turn depend on the normal vibrations of the molecule. Thus the way in which a change in the structure affects the intermediate stages is very complex and it is impossible to create a direct link back from the descriptor to the structure[270,419,420].

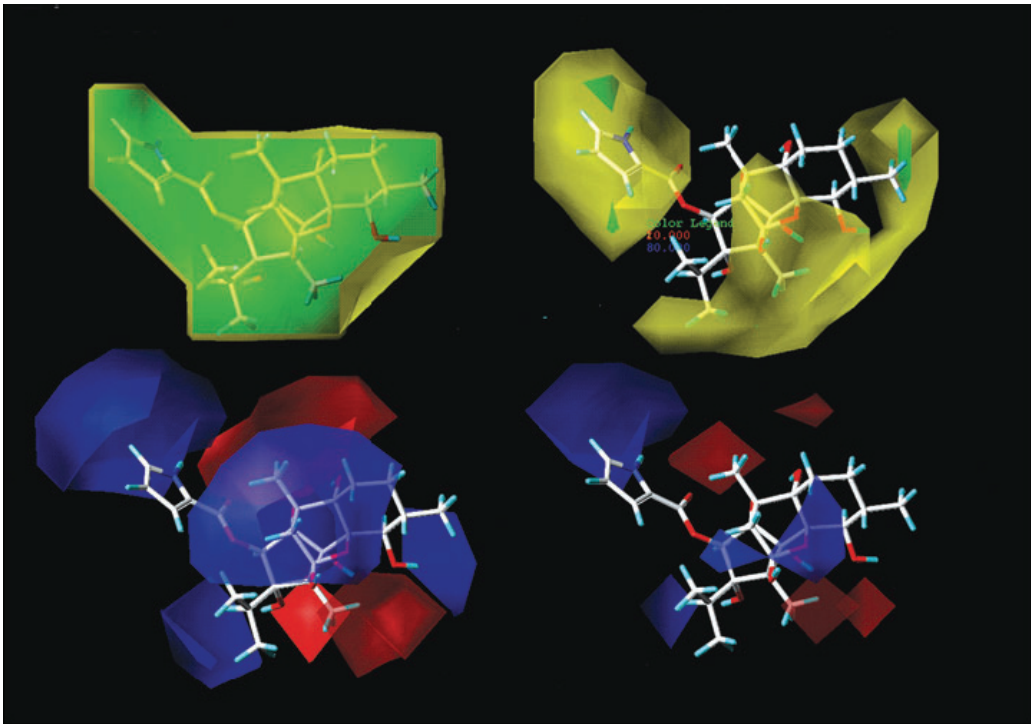


Figure 6. An example of CoMFA interaction contour plots as generated by Sybyl.

3. THE FLUFF-BALL METHOD AND ITS VALIDATION

The primary aim of this study was to create a QSAR technique which is capable of processing large and diverse libraries with minimal user intervention. It should take into account the fuzzy nature of the molecular structures. This is important because often the so-called "toy model of chemistry" becomes dominant and molecules are thought to be a rather static collection of hard spheres. In fact a molecule is a dynamic system in constant motion and the atoms are malleable and do not resemble the hard plastic balls of the toy models. Therefore, the functional forms used to model atoms should also incorporate this malleability and slowly fade away instead of having a clearly defined edge. Also, if one removes the large ordered motions there is still residual motion caused by the unordered thermal vibration of atoms. If there are enough snapshots these small uncertainties in the positions of the atoms can be averaged creating probability distribution for the position of the atom. The volume of an atom defined by a probability cloud can be easily described with fuzzy functions.

As this kind of special QSAR would clearly benefit from a tailor-made superposition technique the pair Flexible Ligand Unified Force Field – Boundless Adaptive Localized Ligand, or FLUFF-BALL, was created. They are a matching pair of superposition and QSAR techniques especially designed to facilitate a rapid analysis of flexible molecule libraries with minimal user intervention. Primary design emphasis has been to balance the computational simplicity necessary for fast screening while ensuring that the FLUFF-BALL remains easily tuneable, allowing the user to import any and all available *a priori* information.

For the FLUFF superposition, the basic insights are centred on ways to gain more information about the receptor to improve the superposition. If the structure of a receptor is known then it can be used for virtual screening[421-425], ligand docking[426-433] or free energy calculations[434-437]. These techniques can be used instead of QSAR, but they can also be used to generate a superposition for a 3D-QSAR analysis. Furthermore, it has been demonstrated that an alignment generated using constraints derived from the receptor model is superior to the standard ligand based alignments[63,64]. Unfortunately, the structures of many receptors are unknown and therefore they cannot be used to derive a set of constraints for a QSAR analysis[438,439]. The FLUFF methodology uses a ligand with high binding affinity as a complementary model of the receptor and thus leverages the implicit information about the binding pocket of the receptor. In FLUFF, the ligand is used as a *template* for superposition, against which all other ligands are aligned. However, that is not the full extent of the information which can be used to improve the performance of QSAR as the dynamic properties of the template have not been utilised. In FLUFF the Gaussian function shape allows the template and ligand to become fuzzy thus in part taking into account the dynamic phenomena caused by small movements and conformational changes which naturally occur in all molecules.

As the available computing power increases, more and more dynamic phenomena are incorporated into QSAR analyses because the model of a rigid key fitting to a rigid lock is not a complete picture of the binding process as both the ligand and receptor are flexible and in constant motion[95,299,440]. Because of the dynamic nature of the binding, one can assume that the conformational space of a high affinity ligand, or some subset of it, is complementary to the confor-

mation space of the receptor. This insight can be used to simplify one of the more arbitrary phases of a 3D-QSAR analysis, namely the selection of active conformation. Usually the geometry optimised structure is selected as the active conformation and for rigid and semi-rigid molecules this is usually a reasonable assumption. On the other hand, as the flexibility of the molecules increases, the chance that the geometry optimisation has found a conformation which is not the active conformation becomes a very real possibility. This is due to the fact that for flexible molecules there are often a considerable number of local optima which have nearly the same energy. Nevertheless, by taking into account the conformational space of the template one can usually drastically reduce the number of possible ligand conformations as the number of conformations which are suitable for both molecules is limited. This reduction of conformational space could be called a *common conformation* approach, and even though this concept itself is quite old, it has not been widely applied to superposition algorithms. In FLUFF the template and the ligand seek together the conformations which yield the best molecular similarity which results in a more realistic picture of the similarity between the ligand and the template.

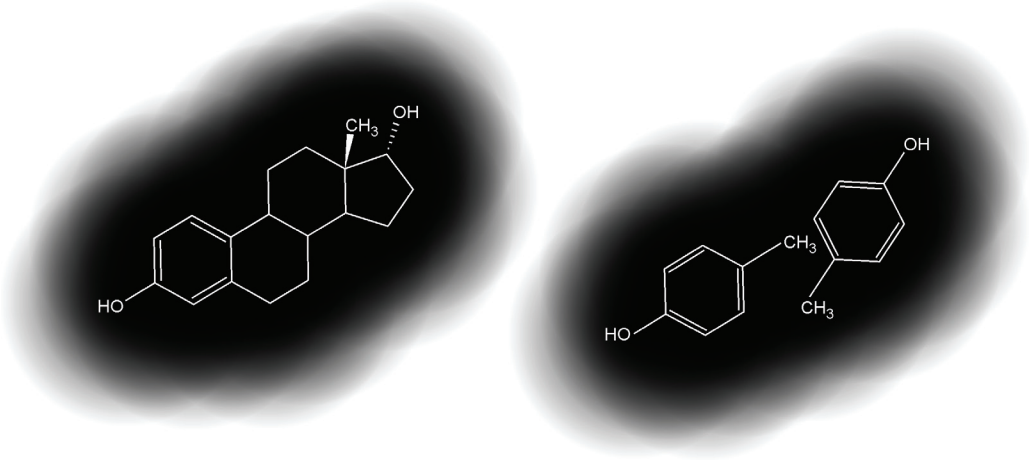


Figure 7. The 17β-estradiol and two *p*-cresol molecules with a set of atom centred circles defined by linear decay density functions.

In superposition one should also consider cases where the template or a ligand consists of several molecules. In order to facilitate alignment of multiple molecules on a template, which could also consist of several molecules we need extend the concept of a molecule. In FLUFF-BALL the standard definition of a molecule as a collection of atoms interconnected with bonds is used to define a *physical molecule*. The ligands and templates are defined as *logical molecules* which may consist of one or more physical molecules, parts of physical molecule(s) or they may even be an arbitrary collection of atoms. The template and ligand(s) used in FLUFF-BALL superposition are logical molecules. For example, in Figure 7 there are three physical molecules: one 17β-estradiol and two molecules of *p*-cresol. However for superposition and QSAR one could

define two logical molecules, one containing the 17β-estradiol and the second consisting from two physical *p*-cresol molecules. It would also be possible to define all three physical molecules into a single logical molecule, or one could also define only the aromatic A-ring and phenolic hydroxyl group as a logical molecule.

For full utilisation of a flexible superposition based on the best common conformation paradigm an accompanying QSAR method is needed as many standard 3D-QSAR -techniques are dependent on a global coordinate system. This means that all molecules must be aligned to the same spatial co-ordinates. In the flexible superposition the demand for a uniform global positioning can not be met. Moreover, if the superposition is performed using a fully flexible template, minor changes may also occur in the conformation of the template whereby the template is no longer a reliable global anchor. Therefore it is most beneficial if the QSAR technique would also use a local grid tied to the centres of the template atoms. One could circumvent the superposition problem altogether by using an alignment-free technique[107,198,267,269,270,272,441-444] but unfortunately they are usually "black-boxes", meaning that it is very difficult or even impossible to back-project the results of the QSAR onto the original molecules and deduce the features which affect the activity of a compound. This severely limits their usefulness in rational drug design[445], although they may be very useful for predictive purposes.

Also, if the QSAR model is built using the implicit information from the high affinity ligand, one should take into account the fact that beyond the volume of the ligand the reliability of the model will decrease as the amount of information available about the shape of the binding pocket decreases and therefore the resolution of the QSAR analysis should also decrease. Also, if the dynamic nature of affinity is taken into account one should use either a multiple copies of the ligand or in some other way include the fuzziness of the ligand structure into the QSAR analysis[440]. The current 3D-QSAR techniques do not fully incorporate this fuzzy nature and the ensemble systems, often called 4D or 5D QSARs[95,440], are in many cases computationally difficult and therefore unsuitable for mass processing of large molecular libraries[446]. Long side-chains are particularly problematic as small changes in the bond and torsion angles at root of the chain can lead to major changes in the position of other end. So in many cases even very similar side-chains form a large fan-like structure which, in conjunction with a dense 3D-QSAR grid, will reduce the accuracy of the model. Therefore it would be better to deal with the problem pro-actively and include the fuzziness into the QSAR analysis from the start. For these reasons the Boundless Adaptive Localized Ligand, or BALL, was created to complement the FLUFF superposition, though it is also capable of processing molecule sets aligned with any other technique.

3.1 Flexible Ligand Unified Force Field (FLUFF)

As stated earlier, the FLUFF superposition methodology aims to better incorporate the dynamic and fuzzy nature of molecules. The superposition algorithm is, in essence, a specialised force field based on a modified Merck Molecular Force Field (MMFF94)[447-451]. This offers many advantages, including the fact that FLUFF can utilise well-documented and tested computational techniques available in the standard molecular mechanics force fields. Also, as a force field, FLUFF can be very easily tuned by adjusting of the superposition parameters for each atom type. It is also possible to include repulsive terms in order to incorporate *not-like-that* type of "negative" superposition rules. The superposition score is expressed as the total energy of the model and therefore the actual superimposition is usually accomplished by performing a geometry optimisation using the superimposition force field. Alternatively, a molecular dynamics (MD) or Monte-Carlo search (MC) can be utilised.

The energy equation (eq. 43) of MMFF94 can be divided into two separate components describing bonded (E_B) and non-bonded (E_{NB}) interactions as follows (eq. 44):

$$E_{MMFF94} = E_B + E_{NB} \tag{43}$$

$$\begin{aligned} E_B &= \sum EB_{ij} + \sum EA_{ijk} + \sum EBA_{ijk} + \sum EOOP_{ijkl} + \sum ET_{ijkl} \\ E_{NB} &= \sum EvdW_{ij} + \sum EQ_{ij} \end{aligned} \tag{44}$$

The bonded term consist of bond length (*EB*), bond angle (*EA*), combined bond angle and stretch (*EBA*), out-of-plain (*EOOP*) and torsion (*ET*) contributions, whereas the non-bonded term has only van der Waals (*EvdW*) and electrostatic contributions (*EQ*). In the FLUFF superposition the normal MMFF94 energy terms are preserved only within a logical molecule and between logical molecules the non-bonded interactions are suppressed thus rendering the logical molecules "invisible" to each other. This means that different logical molecules do not interact and can pass straight through each other. In addition to the bonded interactions and the non-bonded interaction terms within the logical molecule, an E_{SP}-term is generated to describe the similarity of van der Waals (*ESvdw*) and electrostatic field (*ESeel*) of the ligand and the template (eq. 45).

$$E_{SP} = \sum ESvdw_{ij} + \sum ESeel_{ij} \tag{45}$$

Early on it was decided that the functions $ESvdw_{ij}$ and $ESeel_{ij}$ should depend only on the positions, or rather distance between atoms, their MMFF94 types and, in the case of $ESeel_{ij}$, the charges of atoms. These variables were selected because the information is readily available and is consistent with the superposition force field design principles. However, in order to enable the user input of *a priori* information about the relative importance of different molecular features, a scaling factors for $ESvdw_{ij}$ and $ESeel_{ij}$ terms were included.

The actual functions representing the $ESvdw_{ij}$ and $ESeel_{ij}$ terms can be chosen arbitrarily as long as they do not contain asymptotic points, and they have an unambiguous derivative at all points. On the other hand, for the purposes of superposition the function should be as smooth as possible and in order to maximise the convergence rate of the optimisation the second order derivative should be as close to a constant as possible. As the van der Waals surfaces of atoms are actually rather soft and deformable, the superposition functions should also reflect this property and allow for a small amount of fuzziness in the exact position of atoms. In other words the function should have an area of very low gradient as the distance between atoms is nearing zero. Also, it would be beneficial if the function would have a large area of low bias as the distance increases. This enables the template to exert a small force even to the ligand atoms which are very far away, but at the same time this bias is small enough that it does not affect the local superposition of the conformational freedom of the ligand.

When considering the requirements listed above it becomes evident that a sigmoidal shape generated by the Gaussian type function $\left(e^{ax^n}\right)$ is almost optimal for energy term. These Gaussian type functions are widely used in quantum chemistry to describe the electron orbitals [452-454] but they are also used in several superposition algorithms[79,81,455-457]. Also if the exponent *n* of the Gaussian function is even, the derivate of the energy term is separable to three sub-terms which indicate the gradient for x, y and z axes of the Cartesian space, thus further expediting the optimisation process. The Gaussian functions are also very useful as the effective range of the function can be changed easily by manipulating the constant *a* without affecting the maximum energy contribution. If the constant *a* is very large the effective range of the function will be very large and the "image" of the molecule on the energy landscape becomes very diffuse and loses a considerable amount of information. In other words, it is possible to easily control the level of detail available to the superposition algorithm and, for example, to perform gradient runs. One such application would be an optimisation run started with energy functions set to almost infinite range, thus reducing all molecules to featureless spheres leading to a rapid convergence as the optimisation of this simplified system is relatively easy. After convergence the range of the functions can be reduced slightly which increases the amount of structural information available to the superposition, but the change for the previous system is very small and the optimisation once again convergences rapidly. When the cycle of optimisation and parameter manipulation is performed iteratively with sufficiently small steps it is possible to generate superposition in a semi-automated fashion without major user intervention.

Due to the FLUFF's handling of non-bonded interactions, there are some guidelines which should be adhered to when constructing template and ligand(s): First and foremost, an atom can belong to only one logical molecule at a time. It is also unadvisable to define two logical molecules within one physical molecule as the non-bonded interactions would be disrupted but the bonded interactions would be maintained. One can include only a part of a physical molecule in a logical molecule and left a part of it unassigned, in which case the non-bonded interactions between the assigned and unassigned part of the molecule will be disrupted, but it will not interfere with the FLUFF superposition. On the other hand, several physical molecules can be assigned to the same logical molecule without any detrimental effects.

FLUFF can be used to perform a pair-wise superposition in which the set of ligands is divided into a set of separate superposition tasks each containing the template and one ligand. This enables easy multitasking and will allow maximal flexibility in the superposition. As FLUFF seeks the best common conformation for a template ligand pair, the whole set of ligands is not necessarily correctly aligned in the global 3D space, even though each pair is properly aligned. It would be possible to align the templates of each pair to achieve a global superposition but as the template is also allowed to deform the accuracy of such a global alignment can not be guaranteed.

If a global alignment is desired, it is best to simultaneously optimise the whole set of ligands against the template molecule whereby one can utilise the flexible FLUFF superposition while at the same time maintaining a global 3D alignment. This will naturally mean that the template and the whole set of ligands will seek their best common conformation which leads to a considerable loss in the degrees of freedom in the superposition, but for many cases this will not adversely affect the performance of the whole superposition.

However, the fully flexible superposition is not necessarily optimal for all cases and thus the current implementation of the FLUFF superposition algorithm provides three main variants, **FIX**, **MIX** and **FLEX**, which differ in their level of flexibility. The **FIX** superposition uses a rigid superposition algorithm, meaning that a rigid ligand is superimposed on a rigid stationary template using FLUFF force-field and standard energy minimisation procedure. In the **MIX** variant, as the name suggest, a mixture of rigid and flexible superposition is performed by using a fully flexible ligands and a rigid stationary template. In the **FLEX** set both the template and ligand are fully flexible and can adapt their conformations according to the FLUFF field and seek the best common conformation.

3.1.1 The ESvdw term

The following functional form is used for the ESvdw term:

$$ESvdw_{ij} = S_{vdw_i} S_{vdw_j} C_{vdw1} e^{C_{vdw2} r_{ij}^{n_{vdw}}} \tag{46}$$

$$ESvdw'_{ij} = S_{vdw_i} S_{vdw_j} C_{vdw1} e^{C_{vdw2} r_{ij}^{n_{vdw}}} C_{vdw2} n r^{n_{vdw}-1} \tag{47}$$

where $\mathbf{Svdw_i}$ and $\mathbf{Svdw_j}$ are the scaling factors of the atoms i and j (usually $S_i = S_j = 1$), $\mathbf{C_{vdw1}}$ and $\mathbf{C_{vdw2}}$ are constants defined for interaction between atoms of type i and j (C_{vdw1} and C_{vdw2} <0), $\mathbf{r_{ij}}$ is the distance between atoms i and j and $\mathbf{n_{vdw}}$ is the exponent defined for interaction between atoms of type i and j (usually n_{vdw}=2).

When selecting field constants C_{vdw1} and C_{vdw2} some care must be taken to ensure that the minimum occurs at the centres of the atoms, as too wide a fitting function can lead to a set of erroneous minima to be generated between the atoms. The phenomenon of false minima can be seen especially clearly when superimposing two benzene rings. If the constant C_{vdw2} allows the

fitting field to span a too great distance a minimum is created between the template atoms, and in the resulting fit ligand atoms are located exactly on the middle point between two template atoms.

3.1.2 The ESeel term

The ESeel term is analogous to the ESvdw but it uses the partial charges of atoms as additional scaling factors.

$$ESeel_{ij} = S_{eel_i} S_{eel_j} q_i q_j C_{eel1} e^{C_{eel2} r_{ij}^{n_{eel}}} \tag{48}$$

$$ESeel'_{ij} = S_{eel_i} S_{eel_j} q_i q_j C_{eel1} e^{C_{eel2} r_{ij}^{n_{eel}}} C_{eel2} n r_{ij}^{n_{eel}-1} \tag{49}$$

where $\mathbf{Seel_i}$, $\mathbf{Sell_j}$, $\mathbf{C_{eel1}}$, $\mathbf{C_{eel2}}$, $\mathbf{r_{ij}}$ and $\mathbf{n_{eel}}$ are analogous to the definitions presented for the ES_{vdw} term (eq. 46) and $\mathbf{q_i}$ and $\mathbf{q_j}$ are the partial charges of atoms i and j.

This function exhibits an attractive behaviour in the case of two similar charges and a repulsive force for two opposite charges. This can, at times, hamper efficient superposition, particularly in case of highly charged side chains, which are usually also flexible. In order to resolve the problem, a cutting option was included in the superposition algorithm, which removes all repulsive interactions by setting all positive energy terms to zero, but leaves all negative terms, i.e. attractive forces, unchanged. Also values of the $ESeel_{ij}$ function are highly dependent on the charges present in the molecules, which in part complicate balancing the strength of this function. In case of the ESvdw the per-atom strength of the function remains quite stable from ligand to ligand, but in case of ESeel it may occasionally be necessary to adjust the field strength by adjusting the constants C_{eel1} and C_{eel2}.

3.2 Boundless Adaptive Localised Ligand (BALL)

As outlined earlier the Boundless Adaptive Localised Ligand (BALL) technique uses an internal co-ordinate system tied to the template. The grid vertices are placed at the atomic centres of the template molecule, thus rendering the internal co-ordinates immune to global translations and rotations. Also, minor changes in the template conformation do not necessarily have major adverse effect on the accuracy of the model as anchor points of the local grid are tied to the template and transform with the template. In the case of extremely flexible molecules the changes in the template conformation cause a considerable cumulative error and reduce the accuracy of the QSAR model. In these cases the template can be locked and the superposition can be performed using only a flexible ligand. When flexible side chains are connected to a rigid or semi-rigid body, the conformational changes are usually minor and a fully flexible superposition may be utilised.

The atom centred localised grid can be interpreted as an extreme form of variable selection as the grid is extremely sparse and covers only the volume of the template. Due to limitations imposed by the sparse grid the BALL field must use a separate terms to describe the parts of the ligand that are outside of the template volume. This “residual field” is projected as a wave front and subsequently allocated onto the template atoms in such manner that the terms generated are directional but they become increasingly fuzzier the further away for the ligand one moves. Thus long side-chains, whose position is usually very poorly defined, affect the QSAR descriptor but their exact orientation is not crucial. This enables the BALL descriptor to take into account the extra steric bulk present in the ligand while at the same time more accurately describing the flexibility implied by the induced fit model of affinity. The fuzzy nature of the BALL descriptor is further enhanced by the use of volume integration and spheres defined by Gaussian density functions to describe atoms. In its current implementation the BALL, like many other 3D-QSAR methods, uses only the van der Waals and electrostatic fields to describe the molecules. However, the soft spheres defined by density functions which are used to compute van der Waals field can be used to represent any and all atomic properties, such as lipophilicity, polarizability or electron donor/acceptor properties.

After the evaluation of the BALL descriptors the actual QSAR model may be built using arbitrary statistical analysis methodology. In order to further analyse the observed structure-activity relationships it is possible to back-project the QSAR model onto the template molecule from which the descriptors were evaluated. The very sparse BALL grid tied to the centres of template atoms is not directly suitable for creation of (interaction energy) contour maps. On the other hand, the BALL grid is linked directly to the template structure and therefore it can be used to directly estimate the importance of different sites in the template molecule. It is also possible to use a suitable decay function to create a dense CoMFA style grid from the sparse BALL grid and then create standard contour maps. Though one should note that beyond the template structure the BALL descriptors, and therefore also the BALL results, will become fuzzy. After a while the BALL results will average out and the contour map will become meaningless.

3.2.1 van der Waals terms

The common volume, CV, of two atoms A and B defined by a Gaussian primitive (GTF) density function is

$$CV_{AB} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C_{A1}e^{C_{A2}r_{AB}^2} \cdot C_{B1}e^{C_{B2}r_{AB}^2}\,dxdydz \tag{50}$$

where $\mathbf{C_{A1}}$, $\mathbf{C_{B1}}$, $\mathbf{C_{A2}}$, $\mathbf{C_{B2}}$ are constants and $\mathbf{r_a}$ and $\mathbf{r_b}$ are the distances from the centres of the atoms. According to the Gaussian product rule, the product of two Gaussian primitives is a new Gaussian primitive. Thus the product of two primitives is

$$\begin{aligned} CV_{AB} &= C_{A1}C_{B1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{(C_{A2}+C_{B2})r_{AB}^2}\,dxdydz \\ CV_{AB} &= C_{A1}C_{B1}\cdot e^{\left(\frac{C_{A2}C_{B2}r_{AB}^2}{C_{A2}+C_{B2}}\right)}\left(\frac{\pi}{-(C_{A2}+C_{B2})}\right)\sqrt{\frac{\pi}{-(C_{A2}+C_{B2})}} \end{aligned} \tag{51}$$

where $\mathbf{r_{AB}}$ is the distance between the centres of the atoms and the constants $\mathbf{C_{A2}}$ and $\mathbf{C_{B2}}$ <0.

All higher order intersections can be returned to the basic intersection of two Gaussian primitives. For example the common volume of atoms A, B and C (see Figure 8) can be evaluated by first generating the Gaussian function representing the common volume of atoms A and B and then by computing the common volume of this new function and atom C. Let us also examine a case where we are interested in the common volume of atom A in respect to atoms B and C. First we take the common volume of A and B and the common volume of A and C. In addition a correction term is needed to take into account the common volume of A, B and C, and thus $CV_{ABC} = A\cap B + A\cap C - A\cap B\cap C$. In a general case, the intersection terms which have an even number of atoms, increase the common volume and the terms which have an odd number of atoms decrease the common volume.

The volume of atom self-overlap can be computed by following equation:

$$CV_{AA} = C_{A1}^2 \cdot e^{\left(\frac{C_{A2}^2 r_{AA}^2}{2C_{A2}}\right)}\left(\frac{\pi}{-2C_{A2}}\right)\sqrt{\frac{\pi}{-2C_{A2}}} \tag{52}$$

where $\mathbf{r_{AA}}$ is always 0 and the constant $\mathbf{C_{A2}}$ <0. This can be further reduced to

$$CV_{AA} = C_{A1}^2\left(\frac{\pi}{-2C_{A2}}\right)\sqrt{\frac{\pi}{-2C_{A2}}} \tag{53}$$

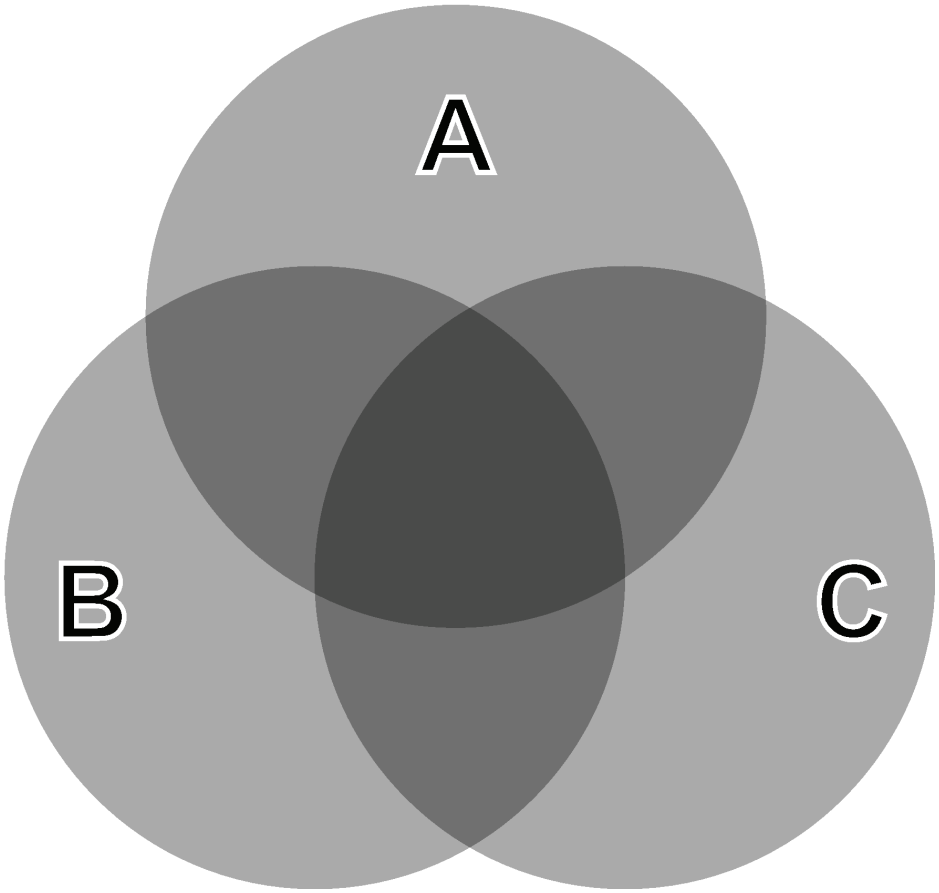


Figure 8. Schematic representation of three atom overlap.

As the volume given by the self-overlap formula (eq. 53) is different to the volume given by the volume integration, it becomes evident that the volume calculations are comparable only within the same order of intersection. Therefore it is necessary to create a conversion formula with which it is possible to generate comparable volumes of intersections. As the first order can also be computed using self-overlap formula to yield the second order equivalent of the atoms volume, the natural choice is to take the second order of intersection as the base level to which all the other intersections are converted. Approximation of a common volume term of the order n-1 is computed from the term of the order n as demonstrated by equation (eq. 54).

$$CV_I = \frac{\sum CV_{n-1}}{n} \cdot \left(\frac{CV_n}{\left(\frac{\sum CV_{n2}}{n} \right)} \right) \tag{54}$$

in which **n** is the order of the intersection, $\mathbf{CV_{n-1}}$ is the intersection of the order n-1 generated by excluding one atom at a time from the original set, $\mathbf{CV_n}$ is the n^{th} order intersection and $\mathbf{CV_{n2}}$ is the pseudo n^{th} order intersections generated from CV_{n-1} intersections by in turn including one atom twice in the intersection. This formula enables the conversion of the 3rd and 4th order intersections to 2nd order with a reasonable accuracy. In test runs it was observed that intersections higher than 3rd order did not improve the accuracy of the model. Therefore the BALL algorithm only evaluates the 1st, 2nd and 3rd order intersections.

In evaluating BALL van der Waals terms the template atoms own volume is computed as a self-overlap by (eq. 53). Then the common volume of the template atom and ligand atoms is computed as an intersection of template atom A and ligand atoms L_1-L_n using (eq. 51) and (eq. 54). The residual volume of the ligand atom means the volume of the atom not covered by the template atoms and it is evaluated in a similar manner as the "free volume" of the template atom, but now the roles of the ligand and the template are inversed. This "residual volume" is allocated to the template atoms by separately evaluating (eq. 55) for each template atom.

$$F_{res} = F_{diff} e^{C_2 r_{tl}} \tag{55}$$

where $\mathbf{F_{res}}$ is the residual field allocated to template atom t, $\mathbf{F_{diff}}$ is the field difference at ligand atom l, $\mathbf{C_2}$ is a constant and $\mathbf{r_{tl}}$ is the distance between the template atom t and the ligand atom l.

To summarise, the following three van der Waals parameters are evaluated for each template atom in order to generate the QSAR descriptor:

1) Template atoms own volume computed with self-overlap (eq. 53).
2) The common volume of the template atom and ligand atoms.
3) The residual volume of the ligand atoms allocated with equation (eq. 55).

3.2.2 Electrostatic terms

In BALL the electrostatic potential V is used to describe the electrostatic similarity of the template and ligand. In general, the electrostatic potential V cast by charge **Q** at point **p** is defined by the following equation:

$$V = \frac{1}{4\pi\varepsilon_0} \cdot \frac{Q}{r_p} \tag{56}$$

where $\boldsymbol{\varepsilon_0}$ is permeability of vacuum ($8.85419 \cdot 10^{-12}$ F/m), **Q** is charge, $\mathbf{r_p}$ is the distance between the charge and point p. However, when computing the QSAR descriptor, the constant term can be omitted, and thus reducing the equation to

$$V = \frac{Q}{r_p} \tag{57}$$

which unfortunately has one asymptotic point at $r_p = 0$ and it is therefore necessary to introduce a limiting factor which prevents the r from reaching zero. In this case a natural limiting factor is the van der Waals radius of the atom used as the point of origin. Thus in BALL the electrostatic potential cast by atom **a** at the centre of atom **b** is evaluated using equation 58 and the field projected by an arbitrary group of atoms **j** at the centre of atom **a** by equation 59

.

$$Sq_{ab} = \begin{cases} \left(\frac{q_b}{r_{ab}}\right) & r_{ab} \geq r_a \\ \left(\frac{q_b}{r_a}\right) & r_{ab} < r_a \end{cases} \tag{58}$$

where $\mathbf{q_b}$ is the charge of atom b, $\mathbf{r_{ab}}$ is the distance between the atoms a and b, $\mathbf{r_a}$ is the van der Waals radius of atom a.

$$Sq_{aj} = \sum_{b=j_1}^{b=j_n} Sq_{ab} \tag{59}$$

Electrostatic terms of the BALL descriptor are obtained by computing the electrostatic field of the template at the centre of a template atom with (eq. 59). Then the difference of electrostatic potential is obtained by evaluating the field projected by ligand with (eq. 59) and computing the difference. The residual potential of the ligand atom means the difference in electrostatic potential projected by the ligand and by the template at the centre of the ligand atom and it is evaluated in a similar manner as the field difference terms of the template but now inversing the roles of the ligand and the template. This potential is then allocated to the nearest template atoms with (eq. 55) as was done with the van der Waals terms.

To summarise, the following three electrostatic parameters are evaluated for each template atom in order to generate the QSAR descriptor:

1) The electrostatic potential projected at the centre of the atom by the template molecule (eq. 59).
2) The difference of potential projected by the template and by the ligand at the centre of the atom.
3) The residual electrostatic potential of the ligand atoms allocated with equation (eq. 55).

3.3 Implementation of FLUFF-BALL

The FLUFF algorithm and the BALL were implemented utilising the MMS software, a molecular mechanics program running under Microsoft Windows, originally developed for use with PERCH NMR software (www.perchsolutions.com) at the Department of Chemistry, University of Kuopio. Though FLUFF-BALL is implemented in the framework of MMS, the algorithm is independent, enabling an easy transfer to other software packages and even to a standalone version. The core of the program, and all the novel algorithms described here, were written in ANSI C++ using standard STL libraries in order to ensure that the code is easily portable to other environments. The molecular graphics are generated utilising the OpenGL and GLUT libraries. The user interface is written with MFC classes using the C++ compiler included in the Microsoft Visual Studio .NET 2003.

The majority of the operations necessary to perform a FLUFF-BALL analysis can be found in the `FLUFF-BALL` menu of the MMS software which, along with its four primary submenus, is presented in the Figure 9. Another major centre of functionality is the FLUFF-BALL Manager (Figure 10) which can be accessed by displaying the MMS manager (`Window / Show MMS Manager`) and selecting the appropriate tab.

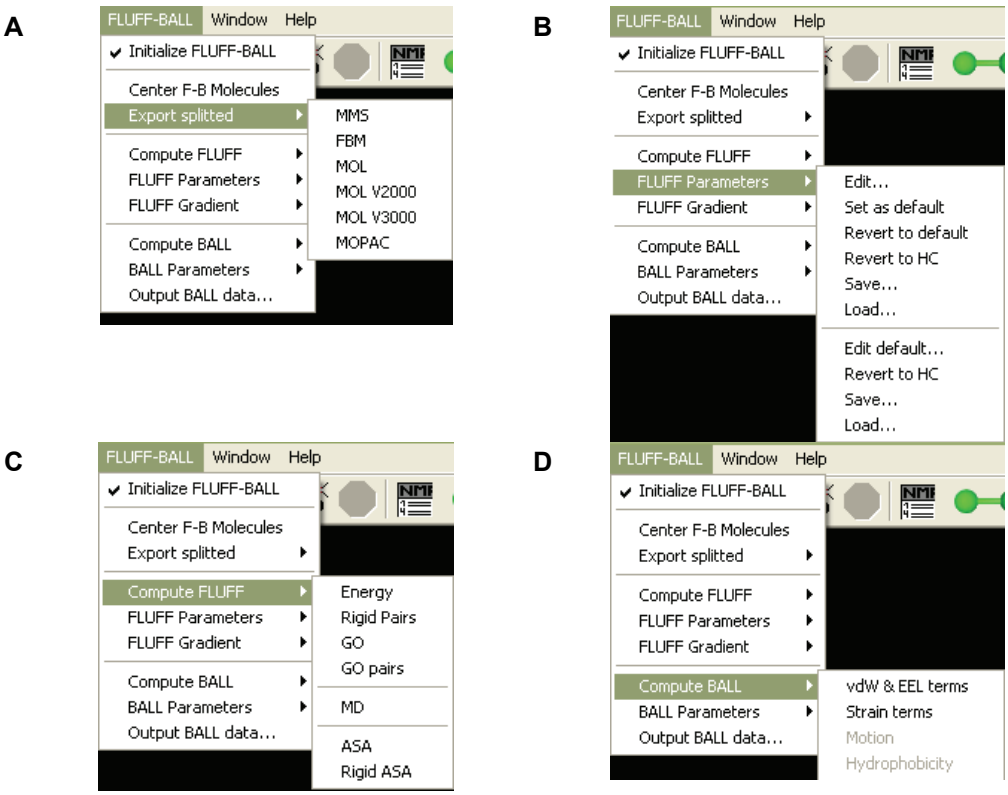


Figure 9 The four primary submenus of the FLUFF-BALL main menu.

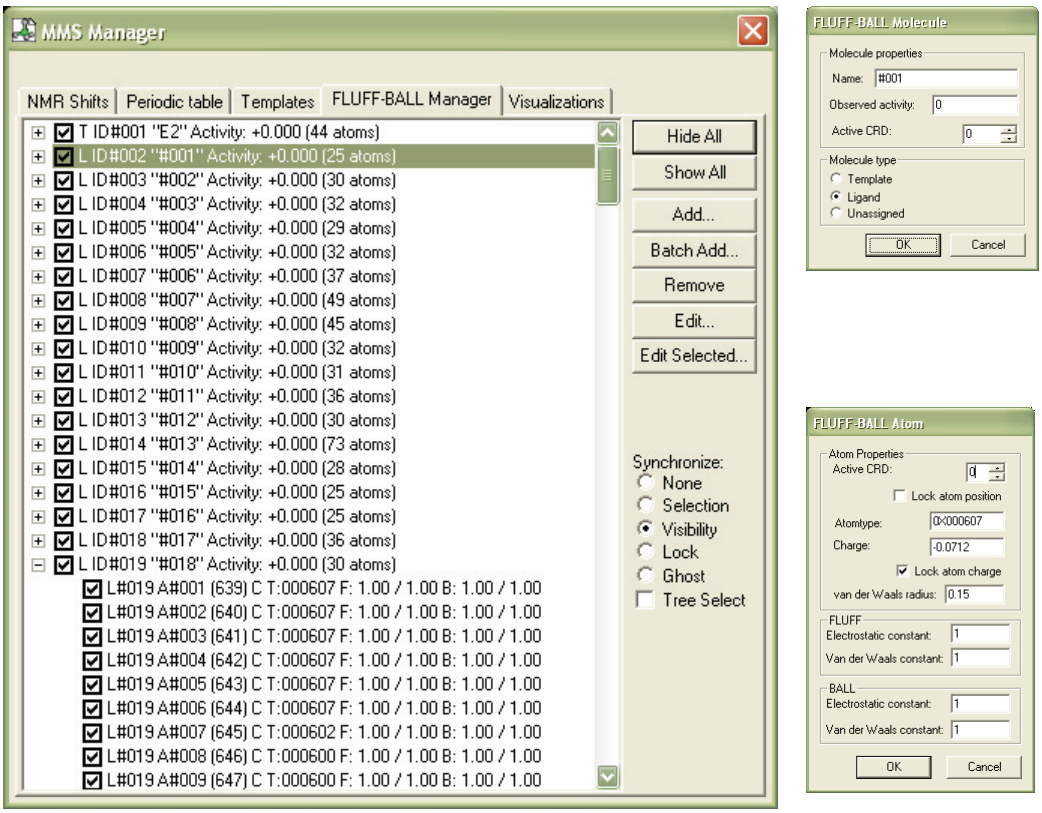


Figure 10 The FLUFF-BALL Manager dialog along with the FLUFF-BALL Molecule and FLUFF-BALL Atom dialogs.

The FB manager is the visual representation of the underlying FLUFF-BALL model and all the functionality required to manage the FB models is present. For a molecule the FB manager displays the type of a molecule (**T**emplate, **L**igand or **U**nassigned), the ID number, name, activity and the number of atoms. The Manager also contains separate items for each atom contained in an FB molecule. The atom information contains the type and ID of the parent molecule, the atom ID, the unique ID of atom in MM model, the element, atom type and the four atom specific scaling factors for FLUFF and BALL terms.

New molecules can be added to the FB manager either by selecting a set of atoms in the model and then pressing the `Add` button or by writing a FBBL file and using the `Batch Add` function. The FBBL file contains a set of lines each of which contains a specification for a new FB molecule (*<TYPE: T/L/U>* "*<NAME>*" *<ACTIVITY>* *<FORMAT: MMS/MOL/HIN>* *<FILENAME>*), the end-of-file tag "END" or a comment beginning with "//". Atom or molecule can be removed by selecting it in the manager and then pressing the `Remove` button. This only removes the atoms/molecules from the FB manager but the MM model will still contains those atoms/molecules. In a similar fashion the atoms and molecules can be edited by selecting them

in the manager and then clicking the `Edit` button. If a large set of atoms is to be edited at the same time one should select them in the model and then press the `Edit Selected` button.

Before any FLUFF-BALL operations can be performed the FLUFF-BALL engine must be initialised. This can be done by clicking the `FLUFF-BALL / Initialize FLUFF-BALL` menu item. The `Export splitted` submenu (Figure 9, submenu A) enables user to export the whole FLUFF-BALL model into a set of files each containing single FB molecule. The primary use of this mode is to export a superimposed set of molecules for further processing in some external program. A template filename of the form *<NAME>.<EXTENSION>* must be supplied and the set of molecules will be exported to a set of files with automatically generated names in the format of *<NAME>_<FB MOLECULENAME>.<EXTENSION>*.

For FLUFF superposition terms six user defined parameters C_{vdw1}, C_{vdw2}, n_{vdw}, C_{eel1}, C_{eel2} and n_{eel} (eq. 46 and 48) are required. These parameters can be edited using the options in `FLUFF-BALL / FLUFF Parameters` menu (Figure 9, submenu B). The dialogs are presented in the Figure 11. In the `FLUFF parameters` dialog the `Is single type` checkbox enables and disables the right column of the Type edit controls and controls whether the parameter is defined for a pair of atom types or a range of atom types. For ease of use the EEL and vdW radius terms are given as half height distances, meaning that the parameters indicate the distance from the centre at which the Gaussian terms has lost half of its value. The six parameters can also be given in a user editable text file, so that it is possible to easily specify a large number of parameters. Each line contains a parameter for a pair of atom types (`S:` *<ATOM TYPE 1> <ATOM TYPE 2>* $<C_{eel1}>$ $<C_{eel2}>$ $<n_{eel}>$ $<C_{vdw1}>$ $<C_{vdw2}>$ $<n_{vdw}>$) or for a range of atom types (`G:` *<ATOM TYPE 1 MIN> - <ATOM TYPE 1 MAX> <ATOM TYPE 2 MIN> - <ATOM TYPE 2 MAX>* $<C_{eel1}>$ $<C_{eel2}>$ $<n_{eel}>$ $<C_{vdw1}>$ $<C_{vdw2}>$ $<n_{vdw}>$), the end-of-file tag "END" or it begins with "//" and is interpreted as comment and discarded.

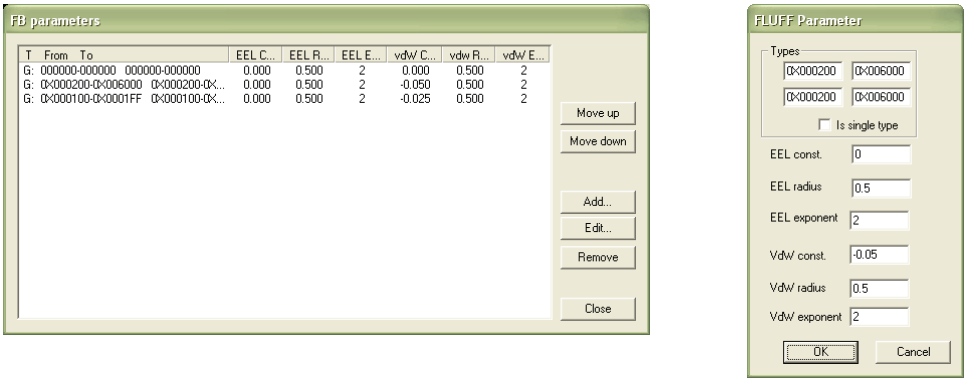


Figure 11 The dialogs allowing the user to edit the FLUFF parameters

The `Compute FLUFF submenu` (Figure 9, submenu C) enables the user to evaluate the current potential energy value of the model (`Energy`), perform a pair-wise rigid, FIX, superposition (`Rigid Pairs`) or the flexible and semi-flexible, FLEX and MIX, superposition for the whole set of molecules at the same time (`GO`) or in a pair-wise computation (`GO Pairs`).

The BALL algorithm requires four user defined parameters. The first two parameters are vdW_constant (vdW_C) and vdW_RadiusIntensity (vdW_RI), which correspond to the C_1 and C_2 constants of the GTF (eq. 50) and control the behaviour of the van der Waals similarity function (eq. 51). The last two, vdW_dispersion (vdW_D) and EEL_dispersion (EEL_D), control the allocation of orphan van der Waals and electrostatic density to the template atoms (eq. 55). They both correspond to the C_2 constant of the GTF. The GUI for editing these parameters can be accessed through the BALL parameters submenu which is identical in appearance to the FLUFF parameters submenu (Figure 9, submenu B). Also, as was the case with FLUFF, the parameters for BALL can also be given in a user editable text file. Each line contains a parameter for a pair of atom type (`S: <ATOM TYPE> <vdW_C> <vdW_RI> <vdW_D> <EEL_D>`) or for a range of atom types (`G: <ATOM TYPE MIN> <ATOM TYPE MAX> <vdW_C> <vdW_RI> <vdW_D> <EEL_D>`), the end-of-file tag "END" or it begins with "//" and is interpreted as comment and discarded.

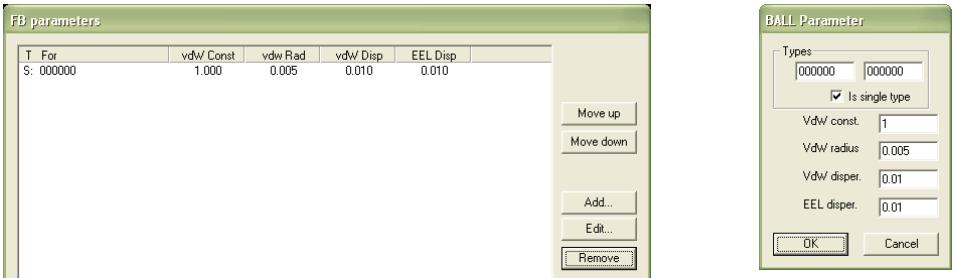


Figure 12 The dialogs allowing the user to edit the BALL parameters

Of the four parameters, the VdW_C is the least significant, as it controls only the intensity of the van der Waals function and not the shape of the function. This parameter is included only for scaling purposes and its value can usually be set to 1. The VdW_RI parameter indicates the intensity of the steric similarity function at the van der Waals radius of the atom. For example the value 0.5 would indicate that the value of this function would be half of what it is at the centre of the atom. Quite naturally this parameter is limited to values between 0 and 1 where zero means that the van der Waals field is immediately quenched and has no effective radius whereas the value one means that the van der Waals field will propagate to infinity and will never dissipate. The dispersion terms vdW_D and EEL_D can range from 0 to infinity where zero means that the dispersion field is constrained to a singular point and no real dispersion is performed and the infinite value results in a uniform dispersion field over the whole model space. Empirical tests showed that the values of the dispersion constants should not exceed 0.5, as the dispersion field will then become too diffuse to have any real value for prediction.

After the parameters have been set the BALL descriptor can be evaluated by using the `Compute BALL` submenu (Figure 9, submenu D), after which the descriptor can be exported using the `Output BALL data` menu item (Figure 13). Usually the most useful output mode is the

matrix formatted file with molecule information as this generates a single file with each row containing first the activity and then the whole BALL descriptor of a single molecule.

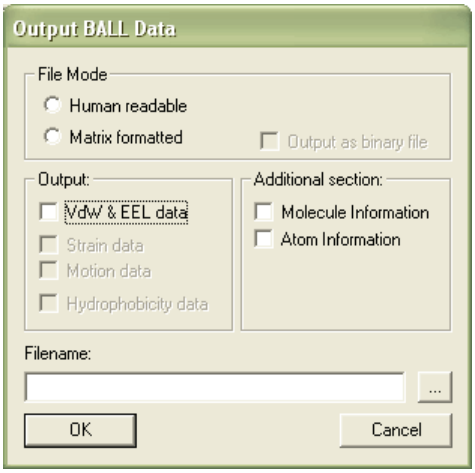


Figure 13 The Output BALL data dialog

3.4 Validation of FLUFF-BALL

The FLUFF-BALL technique was validated by performing an extensive series of tests utilising the following six datasets: The *Cramer* set[14], also referred as *CBG* set, which is a widely used benchmark dataset containing 31 steroids whose binding affinity for the CBG-protein is measured. Many authors[202,265,458] have pointed out that the majority of early works employing the *CBG* dataset contained incorrect structures. Therefore this work utilised a corrected set created for the evaluation of EEVA[271]. The *CBG* set is usually divided into a training set consisting of 21 molecules (**1-21**), and a test set of 10 molecules (**22-31**). The *HALO* set[459] contains 44 halogenated estradiol derivatives whose affinity for the estrogen receptor is measured using receptor binding assay. The *MCF* set[460] contains 42 estradiol-17β analogues for which the K_a values of the receptor-ligand complex and the MCF-7 cell growth response EC_{50} data are available. Both biological activities reported for the *MCF* set were processed as in the original article, thus generating two separate data sets, the *MCF log* K_a and the *MCF* pEC_{50}. In order to gain comparable results molecule **1** (estratriene) was excluded from the MCF data as was also done in the original article. The *PCDD* and *PCDF* sets[210,273] respectively contain 25 halogenated dibenzo-p-dioxin congeners and 34 chlorinated dibenzofuran congeners for which the receptor binding data for cytosolic aromatic hydrocarbon (*Ah*) receptor are known.

All molecules were built using the HYPERCHEM program (version 4.5, http://www.hypercube.com) and subsequently optimised using the AM1 Hamiltonian as implemented in the AMPAC program (version 2.1, QCPE#506). After optimisation the structures were verified by hand to ensure that that all configurations had been correctly assigned and that the molecules had adopted relaxed conformations. For FLUFF superposition an estradiol-17β molecule was imported into *CBG, HALO* and *MCF* datasets and used as a template for FLUFF superposition. In the case of the *PCDD* and *PCDF* sets template molecules dibenzo-p-dioxin and dibenzofuran were imported, respectively.

In order to evaluate the performance of the FLUFF superposition algorithm, three separate sets of superposed molecules were generated using the **FIX**, **FLEX** and **MIX** variants of the FLUFF superposition. During the first tests of the semiautomatic superposition, it was discovered that in several cases the hydrogen atoms of the steroid molecules formed a "local minimum barrier" around the body of the molecule thus interfering with the optimal alignment of the molecules. Therefore the superposition was performed in two separate phases. In the first phase hydrogen atoms were excluded from the FLUFF field, which made them totally transparent to the atoms of a different logical molecule. The other atoms like carbon and oxygen were allowed to see each other. In the second phase, all atoms were included in the FLUFF field. The two-phase optimisation system was automated and is now a standard feature of the FLUFF algorithm. In the superposition of the **MIX** and **FLEX** sets 2500 optimisation steps were used for the first and second phases in order to ensure that the molecules have sufficient time to settle to their optimum conformations. In practice, the analysis of trajectory files showed that after 400-500 steps the energy gradient was less than 10^{-5} kJ/mol and the changes in the conformation were neglible. Only the rough initial superposition was performed manually and the FLUFF algorithm was allowed to find the optimum superposition without any human intervention.

For the superposition results, it must be noted that **MIX** and **FLEX** sets of the *PCDD* and *PCDF* generated almost identical results with very precise alignment of the molecules whereas in the **FIX** sets there was still some difference in the orientation of the substituents. The BALL results of the *PCDD* and *PCDF* sets clearly reflect this as the **FIX** sets generated slightly lower Q^2 values and the optimum components differed from the optimum values of the **FLEX** and **MIX**. For the steroid sets one must note that the steroid backbone is rather rigid and its conformational changes were small. Therefore, the main task was the finding of correct alignment for the side chains. In the **FIX** sets the only criteria for the alignment was the root-mean-square-deviation of the backbone atoms. This resulted in good alignment of the backbones and considerable differences in the side chain conformations. In the **MIX** sets there were some differences in the optimum backbone conformation between the MMFF94 optimised ligand and the AM1 optimised template. This effect was especially clear in the A- and D-rings of the steroid backbone. However, the side chains were generally well aligned. For the **FLEX** sets it was clear that the steroids were assuming a common optimum geometry and the quality of the superposition was more balanced as there were only small errors in both backbone and in side chain alignments. As the FLUFF is a force field-based technique, the superposition can be performed by geometry optimisation, as was done in this work, but any other method for finding minimum energy can also be used.

As the optimum values of the BALL parameters were not known, an extensive optimisation procedure was performed. The value of vdW_C was left at one in all validation runs because earlier testing had indicated that no scaling is needed. For vdW_RI values of 0.950, 0.900, 0.850, 0.800 0.750, 0.700, 0.650, 0.600, 0.550, 0.500, 0.250, 0.125, 0.075, 0.050 and 0.025 were evaluated. For vdW_D and EEL_D values of 0.500, 0.250, 0.125, 0.075 and 0.050 were evaluated. Altogether, this yields a total of 375 unique combinations of parameters. The BALL descriptors were calculated for all datasets using **FIX**, **MIX** and **FLEX** superpositions and aforementioned parameter groups resulting in 1125 different descriptors per dataset.

The statistical computations were performed using SVDPLS regression implemented as MATLAB[461] script. The maximum number of principal components (PCs) was set at 5 for the *CBG* set and to 15 and 11 for the case of *HALO* and *MCF* datasets, respectively. For *PCDD* and *PCDF* sets the maximum number of component was set to 7. After the statistical analyses the optimum BALL models, as indicated by highest Q^2 values, were selected for each dataset using the three different FLUFF superpositions. A summary of these models, showing optimal BALL parameters and some statistical descriptors is presented in Table 3. The Q^2 values yielded by the BALL QSAR clearly indicate that a set of highly predictive QSAR descriptors was successfully generated for each of the datasets.

Table 3. Optimal BALL parameters and statistical descriptors for *CBG*, *HALO*, *MCF*, *PCDD* and *PCDF* datasets.

Set	vdW_RI	vdW_D	EEL_D	Q^2	SDEP	NPC
CBG FIX	0.700	0.500	0.500	0.801	0.552	2
CBG FLEX	0.750	0.120	0.250	0.733	0.678	4
CBG MIX	0.125	0.120	0.120	0.772	0.647	5
HALO FIX	0.850	0.500	0.120	0.659	18.703	13
HALO FLEX	0.750	0.500	0. 500	0.717	17.643	15
HALO MIX	0.800	0.500	0.050	0.643	18.830	12
MCF log K_a FIX	0.800	0.500	0.050	0.339	0.979	4
MCF log K_a FLEX	0.850	0.250	0.050	0.544	0.824	5
MCF log K_a MIX	0.850	0.250	0.050	0.521	0.845	5
MCF pEC_{50} FIX	0.800	0.500	0.050	0.431	1.094	4
MCF pEC_{50} FLEX	0.800	0.500	0.050	0.469	1.057	4
MCF pEC_{50} MIX	0.850	0.500	0.050	0.458	1.067	4
PCDD FIX	0.025	0.050	0.500	0.688	0.883	4
PCDD FLEX	0.850	0.250	0.500	0.728	1.004	7
PCDD MIX	0.850	0.250	0.500	0.728	1.004	7
PCDF FIX	0.075	0.500	0.500	0.727	0.871	7
PCDF FLEX	0.850	0.050	0.500	0.752	1.104	7
PCDF MIX	0.850	0.050	0.500	0.752	1.104	7

When the parameter spreads were evaluated, it became evident that the main parameter affecting the predictive power is the vdW_RI, while vdW_D had a lesser effect. The impact of the EEL_D was increased if the structure contained charged atoms such as halogens but in many cases the charge dispersion did not affect the Q^2 value of the model. The optimum parameters for the *Fix* and *Flex* sets of the standard benchmark were similar whereas the *Mix* set was markedly different from the other two sets (Table 3). All FLUFF-BALL descriptors produced a valid model with high Q^2 values (0.626-0.801). In the *HALO* and *MCF* sets and the *Flex* and *Mix* cases of the *PCDD* and *PCDF* the optimal parameters are located in a small area around VdW_RI of 0.800. The VdW_D and EEL_D parameters have a higher variance but overall they have a lesser impact on the Q^2 value. All in all the BALL parameter optimisation can be focused on the area of VdW_RI 0.700-0.900 and if the two dispersion parameters are included in the optimisation this means that 125 unique sets are to be evaluated. With a powerful desktop PC this optimisation can be performed in less than 20 minutes. Furthermore this optimisation can be performed without human intervention so it only demands computer time. But if the optimisation is to be omitted a BALL parameter set of VdW_RI 0.800, VdW_D 0.500 and EEL_D 0.500 should provide a reasonable Q^2 value. However one should note that this optimum is found using high connectivity structures and may not be universally applicable.

For the *CBG* set the results of the standard test (**22-31**) and training (**1-21**) sets are generally used to evaluate the performance of QSAR techniques and therefore LOO Q^2 values of the whole set of 31 molecules are of lesser importance. As can been seen from the LOO results the **MIX** model created the maximum number of components allowed and therefore an additional sets of statistical descriptors were evaluated with the maximum NPC set to 20. For **FIX** and **FLEX** sets no new models with high number of NPCs were found, but for the **MIX** set a new optimal model with 9 PCs was found. The full results of the FLUFF-BALL models created from the standard division of *CBG* dataset are shown in Table 4.

Table 4. Statistical descriptors of the optimal models generated from the standard division of the *CBG* dataset. The values in parenthesis indicate models where the compound M31 was excluded.

	CBG FIX	CBG FLEX	CBG MIX C5	CBG MIX C20
Spress	0.627	0.740	0.686	0.680
Q^2	0.758	0.682	0.726	0.815
NPC	3	4	4	9
SE	0.414	0.399	0.410	0.173
R^2ex	0.141 (0.561)	0.155 (0.710)	0.072 (0.068)	0.425 (0.180)
SDEP	1.009 (0.687)	0.863 (0.502)	0.712 (0.712)	0.481 (0.492)
Pr-R^2	-0.103 (0.534)	0.193 (0.751)	0.451 (0.573)	0.749 (0.761)

The statistical descriptors of **Mix C5** and **Fix** models are very similar while the results given by the **Flex** model fall below the results of **Mix C5** and **Fix** models. The relatively poor performance of **Fix** and **Flex** models can partially be explained by the very poor prediction of compound **M31**'s activity. The best overall prediction results are achieved by the **Mix C20** model. However this model uses high number of components and is therefore not optimal for predictive use.

If one examines the prediction results of the standard *CBG* test set (Table 5), it is obvious that the compounds **M27** and **M31** are systematically predicted to have too high an activity and only the **MIX C20** model predicts the activities of these molecules with a reasonable accuracy. The reason for the anomalous activities of these molecules can be attributed directly to their structure. In the case of molecule **M27** there are several adjacent hydrogen bond -forming groups which bind to each other in a way that the model can not fully imitate, thus creating a marked error in the predicted values. Molecule **M31** has a fluorine in the 9α-position and it is the only molecule in the steroid set that is *endo*-substituted. In fact, Kubinyi[462] has emphasised that the molecules in the standard training set do not cover all structural features found in the test set. In particular, the compound **M31** requires considerable extrapolation and is therefore a poor test molecule as the QSAR models are only reliable in interpolation. Therefore, and also because the prediction results were systematically poor, compound **M31** was excluded from the prediction set and a new set of prediction runs was performed. The results of these runs are shown in parentheses in Table 4.

Table 5. Prediction results for the standard *CBG* test set (**22-31**). The values in parenthesis indicate the difference between the predicted and observed activities.

Compound	Observed	FIX	FLEX	MIX C5	MIX C20
H22	7.512	8.015 (0.503)	7.822 (0.310)	7.270 (-0.242)	7.164 (-0.348)
H23	7.553	8.098 (0.545)	7.614 (0.061)	7.149 (-0.404)	6.853 (-0.700)
M24	6.779	7.707 (0.928)	7.227 (0.448)	7.628 (0.849)	6.541 (0.238)
H25	7.200	7.702 (0.502)	7.829 (0.629)	7.424 (0.224)	7.119 (0.081)
M26	6.114	6.013 (-0.101)	6.399 (0.285)	6.825 (0.711)	6.871 (0.757)
M27	6.247	7.674 (0.162)	7.285 (-0.227)	7.318 (-0.194)	7.069 (-0.443)
H28	7.120	7.663 (0.543)	7.710 (0.590)	7.460 (0.340)	7.561 (0.441)
M29	6.817	7.251 (0.434)	6.926 (0.109)	7.614 (0.797)	6.993 (0.176)
H30	7.688	7.914 (0.226)	7.927 (0.239)	6.983 (-0.705)	7.519 (-0.169)
M31	5.797	8.234 (2.437)	8.073 (2.276)	6.881 (1.084)	6.171 (0.374)

If the compound **M31** is excluded from the prediction set, the statistical descriptors of **Mix C20** and **Flex** models become similar and both models give good predictive results. The very good performance of **FLEX** model, when compared to the **FIX** model, suggests that the conformational adaptation can significantly aid the construction of a QSAR model. The descriptors of **MIX C5** and **FIX** models are also similar but the predictive results of these models are inferior to the results of **MIX C20** and **FLEX** models. In Table 6 the predictive results of the FLUFF-BALL algorithm are compared to 13 other widely used QSAR methods. While the FLUFF-BALL does not yield the best overall result, its performance is nevertheless comparable to those of most previous QSAR methods.

Table 6. Comparison of FLUFF-BALL with other QSAR techniques for the standard *CBG* test set (**22-31**). The values in parentheses indicate models derived after exclusion the compound M31.

Method	R^2ex	SDEP	$Pr\text{-}r^2$
COMPASS	0.16 (0.69)	0.70 (0.34)	0.46 (0.89)
MS-WHIM	0.28 (0.63)	0.66 (0.41)	0.52 (0.83)
PARM	0.33 (0.30)	0.71 (0.74)	0.45 (0.45)
TQSAR	0.16 (0.36)	0.76 (0.56)	0.37 (0.69)
SOMFA	0.20 (0.62)	0.58 (0.36)	0.63 (0.87)
EVA	0.36 (0.34)	0.53 (0.51)	0.69 (0.74)
CoMFA	0.25 (0.75)	0.71 (0.40)	0.45 (0.84)
GRIND	- (0.88)	- (0.26)	- (0.93)
MFTA	0.87 (0.82)	0.30 (0.31)	0.90 (0.90)
COMSA	0.09 (0.41)	0.70 (0.44)	0.47 (0.81)
MEDV	0.45 (0.57)	0.65 (0.59)	0.54 (0.66)
QS-SM	0.36 (0.22)	0.54 (0.49)	0.68 (0.76)
EEVA	0.36 (0.58)	0.58 (0.40)	0.64 (0.85)
FLUFF-BALL FIX	0.14 (0.56)	1.01 (0.69)	-0.10 (0.53)
FLUFF-BALL FLEX	0.16 (0.71)	0.86 (0.50)	0.19 (0.75)
FLUFF-BALL MIX C5	0.07 (0.07)	0.71 (0.71)	0.45 (0.57)
FLUFF-BALL MIX C20	0.43 (0.18)	0.48 (0.49)	0.75 (0.76)

Even though the results of the CBG dataset are most promising, there are still four other datasets, namely *HALO*, *MCF*, *PCDD* and *PCDF*, which should be analysed before any conclusions are drawn about the performance of the FLUFF-BALL methodology. The Q^2 values of the *HALO* set (0.643-0.717) are fully comparable to the ones obtained in the original article (0.566-0.767) using SYBYL field fitting and CoMFA. The *MCF log* K_a models also yielded Q^2 values (0.339-0.544) which are comparable to the CoMFA models obtained using RMS fit of the steroid backbone (0.395-0.583) and also to the results of SEAL fit (0.426-0.597). The *MCF* pEC_{50} set produced slightly lower values (0.431-0.469) and the difference to the CoMFA models obtained using RMS fit of the steroid backbone (0.463-0.624) or the SEAL fit (0.424-0.582) was significant. In general one should note that the *HALO* set generated a high number of components and when the maximum number of allowed components was raised several models generated up to 20 components. However these data were disregarded because such a high number of components indicate a considerable over-fitting. The *PCDD* set yielded Q^2 values (0.688-0.728) that were lower but still comparable to the values reported in the literature[210,265,273,463,464] (0.715-0.862). The *PCDF* set generated slightly better Q^2 results (0.727-0.752) which are also closer to the values reported in the literature[210,273,463,464] (0.742-0.795). In general one should note that the **FLEX** models generated the best Q^2 values closely followed by the **MIX** models and the **FIX** models generated the poorest models. In the case of *PCDD* and *PCDF* the **FLEX** and **MIX**

resulted in almost identical superpositions, but in the **FIX** set the halogen substituents were slightly offset because of the small changes in the optimal backbone conformation.

For each superposition and dataset combination a total 1000 Y-scrambling runs were performed and in all cases the predictive ability was completely lost, thus indicating that the correlation observed with the correct data is not fortuitous. The external validation was performed using the bootstrapping methodology by creating a collection of 2500 random partitions to test/training sets consisting of 10/21, 15/29 and 14/28 molecules for *CBG*, *HALO* and *MCF* datasets, respectively. For PCDD and *PCDF* 5 compounds were separated for the test set leaving 21 and 29 compounds respectively for the training set. The maximum number of components was set at 10 for *HALO* and *MCF* datasets and at 5 for *PCDD* and *PCDF*. The average results and the standard deviations of the 2500 runs are shown in Table 7.

The *HALO* and *MCF* sets, which were known to be computationally difficult, generated a wide spectrum of models as is indicated by the high standard deviations of the statistical descriptors. However, all models were clearly predictive as indicated by positive Pr-R^2 values even though the results of the HALO and MCF sets are not particularly high and the standard deviation is considerable. Especially in the case of the *HALO* this instability most likely stems from the uneven distribution of the observed values. It also seems that for these datasets there is no discernable performance difference between the three FLUFF superposition variants. On the other hand, for the *PCDD* set and **FIX** variant the external validation failed for many cases as indicated by low Pr-R^2 value (0.113) and disproportionally high standard deviation (2.446) while for the *PCDF* set the **FIX** methodology worked rather well. The **FLEX** and **MIX** variants generated models with high average Q^2 values and high predictivity for the *PCDD* and *PCDF* sets as indicated by Pr-R^2 values of 0.438-0.549.

For the *CBG* dataset the **MIX** model gives the best average prediction and the **FLEX** set gives the least significant results. Again the **MIX** set generates the highest amount of components and its optimum number of components is again clearly limited by the maximum of five components. When the same run was performed with the maximum number of components set to 20 **FIX** and **FLEX** sets generated no new models with a high number of optimum NPC. Yet, for the **MIX** set there are 1390 models for which the optimum number of components was higher than five, and the highest number of components generated is 11. Distribution of the optimum number of components in the *Mix* set is shown as a histogram in Figure 14. A most interesting pattern of two separate clusters of optimum components is observed. When the scrambling run is done to the **MIX C20** set predictive ability is lost indicating that the high Q^2 values observed are not caused by chance correlations.

Table 7. Average statistical descriptors of the models generated from 2500 random training and prediction sets generated from *HALO*, *MCF*, *PCDD, PCDF* and *CBG* datasets.

Set	Spress ±SD	Q^2 ±SD	NPC ±SD	SE ±SD	R^2ex ±SD	SDEP ±SD	Pr-R^2 ±SD
HALO FIX	25.435 ±2.992	0.433 ±0.145	9.6 ±1.0	7.074 ±2.960	0.529 ±0.256	20.006 ±6.941	0.279 ±0.505
HALO FLEX	22.978 ±3.511	0.495 ±0.177	9.3 ±1.1	8.064 ±3.140	0.571 ±0.235	19.585 ±7.042	0.285 ±0.655
HALO MIX	23.526 ±2.947	0.482 ±0.166	9.5 ±1.0	11.041 ±3.861	0.528 ±0.231	20.753 ±8.495	0.222 ±0.819
MCF log K_a FIX	1.103 ±0.101	0.208 ±0.141	3.5 ±1.2	0.822 ±0.160	0.235 ±0.156	1.087 ±0.175	0.120 ±0.270
MCF log K_a FLEX	1.030 ±0.109	0.336 ±0.147	4.7 ±1.3	0.723 ±0.125	0.348 ±0.181	0.995 ±0.190	0.251 ±0.325
MCF log K_a MIX	1.066 ±0.111	0.295 ±0.155	4.8 ±1.5	0.742 ±0.143	0.303 ±0.177	1.032 ±0.192	0.199 ±0.294
MCF pEC_{50} FIX	1.183±0. 088	0.375 ±0.127	4.0 ±1.1	0.860 ±0.129	0.398 ±0.264	1.175 ±0.297	0.228 ±0.506
MCF pEC_{50} FLEX	1.172 ±0.102	0.429 ±0.109	5.8 ±1.7	0.703 ±0.141	0.471 ±0.249	1.107 ±0.298	0.264 ±0.589
MCF pEC_{50} MIX	1.890 ±0.100	0.416 ±0.109	6.0 ±1.7	0.703 ±0.133	0.458 ±0.252	1.118 ±0.295	0.231 ±0.590
PCDD FIX	1.009 ±0.150	0.648 ±0.078	4.5 ±0.5	0.507 ±0.121	0.687 ±0.248	1.052 ±0.567	0.113 ±2.446
PCDD FLEX	1.137 ±0.196	0.643 ±0.079	5.0 ±0.2	0.476 ±0.131	0.628 ±0.254	0.966 ±0.281	0.458 ±0.399
PCDD MIX	1.137 ±0.196	0.643 ±0.079	5.0 ±0.2	0.472 ±0.136	0.619 ±0.262	0.973 ±0.295	0.438 ±0.483
PCDF FIX	0.898 ±0.101	0.700 ±0.070	4.6 ±0.7	0.598 ±0.072	0.683 ±0.223	0.787 ±0.198	0.231 ±0.307
PCDF FLEX	0.926 ±0.120	0.657 ±0.086	4.7 ±0.6	0.507 ±0.138	0.637 ±0.223	0.865 ±0.232	0.549 ±0.289
PCDF MIX	0.934 ±0.119	0.657 ±0.085	4.7 ±0.6	0.504 ±0.143	0.623 ±0.232	0.875 ±0.238	0.522 ±0.356
CBG FIX	0.738 ±0.110	0.573 ±0.134	2.4 ±0.7	0.569 ±0.137	0.619 ±0.175	0.709 ±0.170	0.543 ±0.261
CBG FLEX	0.780 ±0.108	0.540 ±0.138	3.0 ±0.9	0.534 ±0.147	0.570 ±0.172	0.753 ±0.153	0.490 ±0.251
CBG MIX C5	0.698 ±0.070	0.664 ±0.084	4.3 ±0.8	0.337 ±0.090	0.695 ±0.149	0.618 ±0.143	0.662 ±0.166
CBG MIX C20	0.672 ±0.084	0.710 ±0.100	5.7 ±1.8	0.261 ±0.126	0.730 ±0.145	0.586 ±0.153	0.694 ±0.162

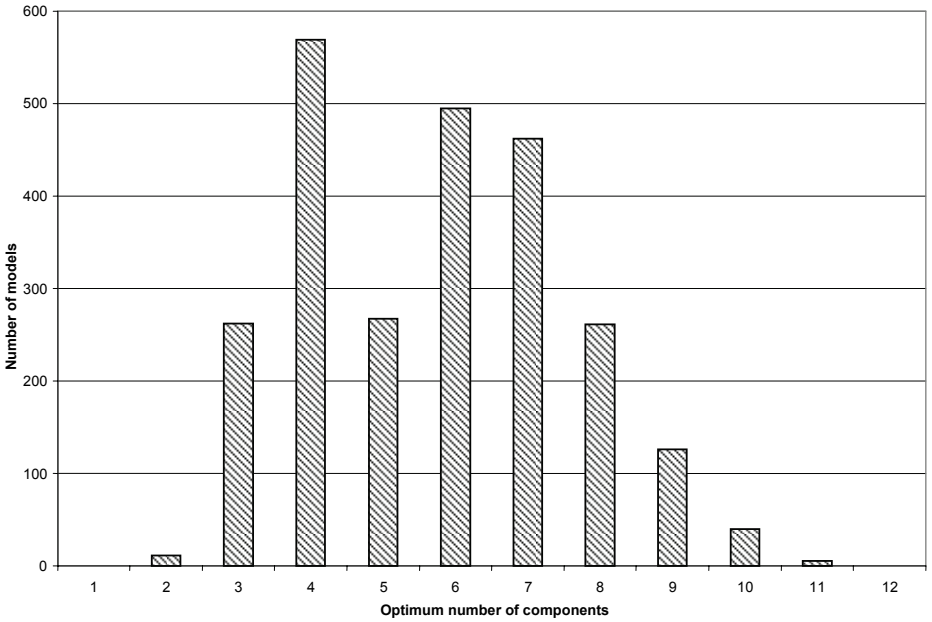


Figure 14. Histogram of the optimum number of components for scrambled Mix C20 models.

3.5 Validating FLUFF-BALL with a large and diverse xenoestrogen dataset

One of the primary design principles of FLUFF-BALL was to create a highly automated superposition and QSAR technique capable of acting as a computational sieve separating the active molecules from a large and diverse molecular library. Therefore, it was deemed necessary to perform an additional validation using much larger and more diverse set than is normally used in the validation of QSAR techniques.

Also, it was decided that the widely used SEAL superposition and CoMFA QSAR techniques should be used to evaluate the same dataset to gain a more reliable benchmark for the relative performance of the FLUFF-BALL methodology. Even though some knowledge of the optimal BALL parameters[465] exists, in this validation work a full optimisation of BALL parameters was performed. This was done in order to evaluate the stability of the BALL model in detail and it should be noted in standard use this kind of procedure would not be necessary. On the other hand, when validating a recently-developed technique, it is vital that the performance limits of the technique are thoroughly evaluated. In contrast, no parameter optimisation was performed for SEAL and CoMFA as they were used in this work as reference techniques, primarily to establish a baseline predictive ability for the xenoestrogen dataset used in this work. Also, both SEAL and CoMFA are well established techniques whose behaviour and optimal parameters have been mapped out over years and numerous applications. On the other hand, FLUFF-BALL is a new technique and its behaviour and optimal parameters are largely unknown at this time.

The term xenoestrogen refers to a chemical compounds which can disturb the natural hormonal balance by binding to the estrogen receptor (ER) in an agonistic or antagonistic fashion[466,467]. The high importance of this class of compounds is due to the highly promiscuous nature of the ER which means there are tens of thousands of molecules, both natural and synthetic, which can bind to the ER and lead to a disruption of the natural hormonal balance. As the hazard these chemicals pose to the environment and to human health has been recognised, they have become the subject of extensive study[466,468,469]. Even though the assay for estrogenic activity is a relatively simple experiment, and there are numerous experimental methods available, the testing of the hundreds of thousands of molecules with in vivo or in vitro techniques for possible estrogenic activity is virtually impossible[467,470-475]. Therefore the benefits of computational screening of xenoestrogens using quantitative structure-activity relationships (QSARs) are obvious[102,467,476,477]. For further information about QSAR models applied to xenoestrogens reader is referred to reviews by Fang et al[476] and Schmieder et al[477]

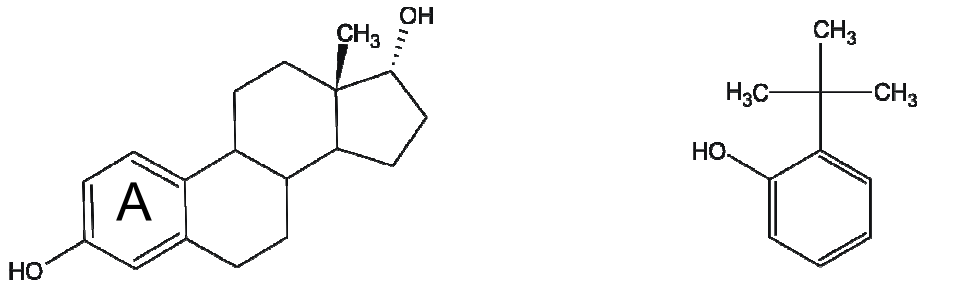
The estrogen binding affinities used in this work were obtained from a freely available standalone version of the endocrine disruptor knowledge base (EDKB, http://edkb.fda.gov) maintained by National Centre for Toxicological Research (NCTR). The EDKB contains about 2000 molecules, many of which do not bind to the ER, rendering the whole of EDKB as such useless as a QSAR benchmark set. Therefore the following selection criteria were used to filter a subset of the EDKB for this work: (1) molecules must have an experimental binding affinity data present, (2) molecules must have detectable binding affinity to ER, and (3) molecules should be small to medium in size. This filtering extracted a subset of 245 molecules containing experi-

mental relative binding affinities (RBA) values for five different estrogen receptors. As some molecules contained experimental values for several receptors there were a total of 374 log RBA values present for 245 molecules. In cases where several RBA values for same receptor were present in the EDKB, the same experimental source was preferred as far as possible. Each estrogen receptor type was treated as a separate dataset yielding five evenly distributed sets of experimental values (Table 8).

Table 8. The five EDKB datasets used.

Receptor	Molecules	Average log RBA (min - max)
Calf	53	0.40 (-2.00 – 2.00)
Human α	61	-0.05 (-2.00 – 2.48)
Human β	61	0.05 (-2.00 – 2.61)
Mouse	69	0.00 (-3.36 – 2.94)
Rat	130	-1.42 (-4.50 – 2.60)

The molecules were built and optimised using the same procedure as in the validation of FLUFF-BALL (see page 71) and subsequently imported to an in-house MMS program (a modified version based on R2004.07, http://www.perchsolutions.com) so that the AM1 optimised co-ordinates and charges were preserved. After importing the molecules were centred according to their centres of mass. Estradiol-17β (E2, Figure 15) was imported as a template molecule, required by the FLUFF and SEAL superposition algorithms. Then the compounds were initially superimposed using a rigid FLUFF superposition on the aromatic ring (marked A in Figure 15) of the template.



Estradiol (E2) **EDKB001 (2-*tert*-butylphenol)**

Figure 15. The structures of the template and the first ligand EKDB001.

For FLUFF superposition the tentatively superimposed set was modified by including or excluding the methyl group (C18, attached to C13) from the estradiol-17β template thus creating **CI** and **CE** sets, respectively. This was done as trial superposition runs indicated that in some cases the methyl group hinders the matching of the backbone atoms. Further sets were generated by exclusion of the hydrogen atoms from the FLUFF field during the superposition, thus creating the **EH** sets. This was done in order to eliminate the barrier effect, which could also hinder the backbone superposition. After the initial superposition without hydrogen atoms, the

IH sets were generated by including the hydrogen atoms in the FLUFF field and performing a full superposition using the **EH** set as an initial guess. For more details the reader is referred to Figure 16 which contains a flow-chart representation of the process used in the creation of the different FLUFF superpositions

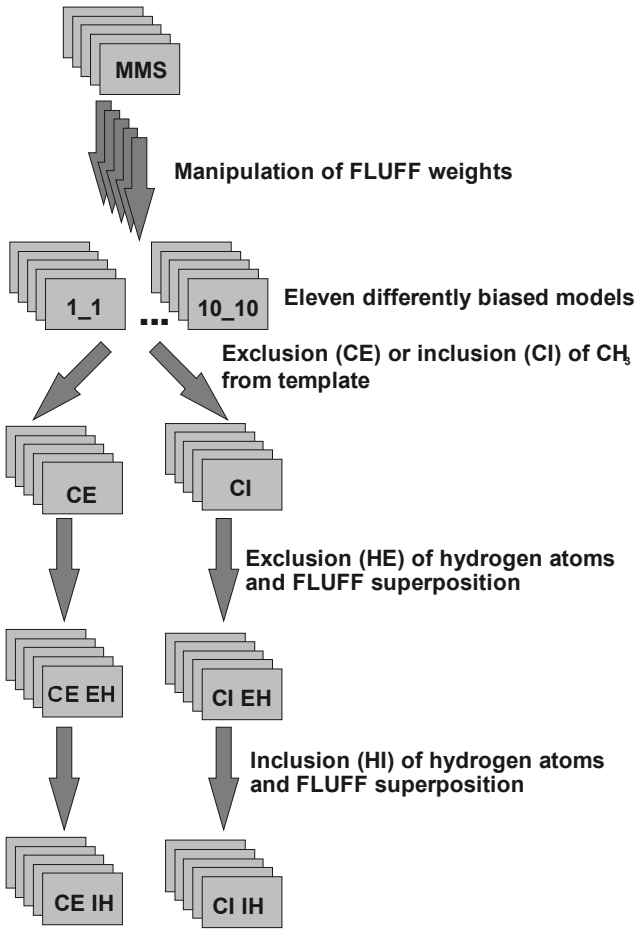


Figure 16. Flowchart of the FLUFF superposition and the creation of the different alignments.

Thus for each of the five experimental datasets a total of 12 different FLUFF models were generated. Nevertheless, it soon became apparent that the EDKB dataset is too diverse to be unambiguously superimposed without any *a priori* information. This can be easily demonstrated using the template and the first ligand EDKB001 (Figure 15). As can be seen there is no unique alignment for the EDKB001, if the only criteria used were the steric and electrostatic properties of the molecules. The EDKB001 can be placed upon the *A* ring of the template in many different orientations, but the aromatic ring of the EDKB001 can also be matched to *B*, *C* or even to the D ring of the template as well. Therefore, additional information about the relative importance of the different molecular features of the template is required. As FLUFF is in essence a

special molecular mechanics force field, the atom types could be used to provide additional information. However, trial runs performed using other datasets[465] have indicated that this kind of selectivity works well only in cases where the molecules are fairly similar. Unfortunately, this is not the case with the EDKB data, and other constraints for guiding the superposition must be used. The current implementation of FLUFF enables the user to assign arbitrary weight factors for the template and ligand atoms and thus increase the amount of information available to the superposition algorithm. In this case several previous QSAR studies[63,220,478-481] suggested that for estradiol-17β the aromatic *A* ring and the hydroxyl group attached to it are important for the biological activity. Therefore the *A* ring of the template was selected as the target for the weight factor modifications and eleven different weight factor combinations were generated (Table 9).

The **1_1** set corresponds to the unbiased superposition and the set **10_10** leads to a superposition where the *A* ring has roughly equal weight as the remaining part of the molecule. Some trials were also done using weight factors up to 25, but they led to models where the *A* ring was overly dominant and caused the ligands to spread out in a fan-like formation. These heavily biased superpositions yielded inferior QSAR models and so they were discarded.

Table 9. The names and weight factors of the eleven FLUFF models used to test the effect of the directed superposition.

Name	A ring OH weight	A ring weight
1_1	1.0	1.0
2_2	2.0	2.0
3_2	3.0	2.0
3_3	3.0	3.0
4_2	4.0	2.0
4_4	4.0	4.0
5_2	5.0	2.0
5_5	5.0	4.0
10_2	10.0	2.0
10_5	10.0	5.0
10_10	10.0	10.0

SEAL superposition was performed by using an in-house SPL script in conjunction with a TRIPOS SPL script (seal.spl as distributed with the SYBYL program, version 6.9.1) and the SEAL program (QCPE #634). In order to import the molecular structures to the SYBYL program[94], the tentative superposition was exported from MMS program as MDL MOL -files, which were converted to the TRIPOS MOL2 -format with the OpenBabel program (version 1.100.2, http://openbabel.sourceforge.net). The MOL2-files were used to generate a SYBYL molecular database, in which the SEAL superposition was performed. All optional parameters were set at the default values present in the TRIPOS script file.

Although some knowledge of the optimal BALL parameters[465] exists, it was decided that the full optimisation procedure should be performed in order to evaluate the stability of the BALL model in detail. In standard use this kind of procedure would not be necessary, but when validating a recently-developed technique, it is vital that the performance limits of the technique are thoroughly evaluated. In contrast, no parameter optimisation was performed for SEAL and

CoMFA as they were used in this work as reference techniques, primarily to establish a baseline predictive ability for the xenoestrogen dataset used in this work. Also, both SEAL and CoMFA are well established techniques whose behaviour and optimal parameters have been mapped out over years and numerous applications. On the other hand, FLUFF-BALL is a new technique and its behaviour and optimal parameters are unknown at this time.

The CoMFA descriptors were evaluated using SYBYL and an in-house SPL script. Using an automatically generated region, standard CoMFA fields containing both steric and electrostatic interactions were generated using a standard distance decay ($1/r^2$) for the computation of dielectric terms, no smoothing, and a 30.0 kcal/mol cut-off with smooth transition for both steric and electrostatic energy terms. For statistical analysis the CoMFA descriptors were exported as text files from SYBYL using an in-house SPL script.

The QSAR models were generated with SVDPLS method and LOO CV using MATLAB[461] scripts. The maximum number of principal components (PC) was set at 15 based on the generally accepted one-quarter rule. For the *RAT* set the maximum number of principal components could be as high as 32 and still conform to the one-quarter rule and therefore additional PLS models with the maximum number of principal components set at 25 and 30 were generated. Some increase in the Q^2 values were observed, but the benefits were negligible (<0.050) and the number of components rose dramatically, being in the range of 25 to 28. Therefore it was judged that the gains made in the Q^2 values did not outweigh the dramatic increase in the number of principal components and these models were discarded.

For each of the five datasets there were originally 49,875 BALL models and 133 CoMFA models. Therefore the results are "distilled" so that for each superposition the optimum BALL model (**FIX**, **FLEX** and **MIX** with **CE** / **CI** and **EH** / **IH** modifications) was selected for further analysis using the maximum Q^2 value. The 133 remaining BALL and CoMFA models were filtered further by selecting the optimum FLUFF weight factors based on the maximum Q^2 value achieved, thus reducing the 132 FLUFF superpositions down to 12. In the summarised results for the five datasets (Tables 3-7) only the reference set generated using SEAL and range of descriptors yielded by the **CE** / **CI** and **EH** / **IH** sets is shown thus reducing the number of FLUFF QSAR models down to 3 for both BALL and CoMFA. This was done as the effect of the modifications was minor.

3.5.1 Effect of Superposition on QSAR

When different FLUFF superpositions are compared, it is evident that no clear optima can be found and therefore it is difficult to discern the optimal FLUFF variant between the **FIX**, **FLEX** and **MIX**. If only the best models of each data set are compared for both BALL and CoMFA, it appears that 3 out of 5 models are generated using the **FIX** method. On the other hand, taking the best 6 of the models for each dataset and making a similar analysis no such trend is observed. If the optimal FLUFF variant is difficult to discern, the optimal weight factors proved to be even more elusive as all available weight factor combinations are present among the optimal models generated for BALL and CoMFA. In fact, the only clear pattern was observed in the CALF dataset, in which 7 out of 12 FLUFF superpositions used the weight parameter set **5_2**.

In general, medium weight factors seem to be favoured, although there are some models that clearly prefer strict constraints. In general, it seems that BALL prefers stricter constraints than CoMFA.

The primary reason for the fact that no clear optima could be found for the FLUFF superposition parameters was that the QSAR models form a large plateau of good predictive ability where the relative differences in the performance of the QSAR are very small. Furthermore, this plateau is dotted with models of higher predictive ability which, however, do not form any clear pattern. It seems that a good QSAR model can be derived from a wide variety of FLUFF superpositions and the optimal models occur as random spikes from the plateau of good performance. This is in agreement with the well-known fact that the 3D-QSAR is highly sensitive to the superposition[5,63,482]. Therefore one must conclude that the choice of FLUFF superposition and the use of *a priori* information in the form of weight factors must be decided on case-by-case basis, and no universal guidelines can be given at this time.

On the other hand, a clear difference in performance can be observed between SEAL and FLUFF (Table 10). Of course, one should bear in mind that the SEAL superposition is only an unoptimised benchmark, but even then the difference between FLUFF and SEAL is significant. This difference could be caused by the force-field nature of the FLUFF technique as it has the ability to provide additional information about the bond patterns and the neighbours of an atom through the use of molecular mechanical atom types that implicitly contain this information. A possible explanation for the superior performance of the FLUFF algorithm is also the fact that it has additional *a priori* information about the relative importance of the features of the template in the form of the user-specified weight factors. The fact that the differences between the **FIX**, **FLEX** and **MIX** variants were smaller than the differences between weight factors suggest that this *a priori* information plays an important role in determining the efficacy of the superposition.

3.5.2 QSAR Results

It seems obvious that the unoptimised reference technique SEAL produces an inferior superposition compared to FLUFF, as indicated by the lower Spress and higher Q^2 values of both FLUFF-BALL and FLUFF-CoMFA (Tables 3-8). In particular, the SEAL-CoMFA combination seems to be particularly problematic as it generates a reasonable model only for the *HUMANB* dataset. The SEAL-BALL combination performs clearly better, but it still produces inferior results when compared with FLUFF. In general, BALL produces better models than CoMFA for both FLUFF and SEAL superpositions with the exception of the RAT data set, for which CoMFA with the FLUFF superposition yielded a slightly better model than BALL. The similar differences in predictive ability that can be observed between the optimum models generated for each dataset also exists in the average results (Table 10), indicating that a real difference exists. Naturally the differences were much more marked between the optimum models.

Table 10. The maximum Q^2 values achieved in LOO CV for all datasets and superposition-QSAR pairs.

	CALF	HUMANA	HUMANB	MOUSE	RAT
FLUFF-BALL	0.824	0.761	0.606	0.611	0.547
FLUFF-CoMFA	0.530	0.407	0.383	0.482	0.673
SEAL-BALL	0.223	0.375	0.410	0.362	0.385
SEAL-CoMFA	0.117	0.163	0.279	0.178	0.157

For each of the EDKB datasets, 1000 Y-scrambling runs were performed and in all cases the predictive ability was completely lost. External validation was performed using the bootstrapping methodology by creating a collection of 2500 random partitions to test/training sets consisting of 18/35, 20/41, 20/41, 23/46 and 43/87 compounds for *CALF*, *HUMANA*, *HUMANB*, *MOUSE* and *RAT,* respectively. As expected, the statistical performance indicators worsened as a result of the bootstrapping, but all models still produced reasonable Q^2 values. The relative performance of the different superpositions changed, but usually the changes were minor and although the order may have changed, usually the same sets can still be found in the top six. However, this is not true for the *MOUSE* dataset where the top three superpositions for BALL were all **FIX** sets which also produced BALL models with unusually low vdW_RI values. When these superpositions were run through the external validation the statistical indicators were significantly lower than those generated from other *MOUSE* sets. As a result three **MIX** sets replaced the **FIX** sets as the top three superpositions for *MOUSE* and BALL.

For the FLUFF superposition the results of external validation for the *HUMANA*, *HUMANB*, *MOUSE* and *RAT* datasets were very similar to the results obtained from the internal validation. BALL still produced slightly better models, except for the RAT dataset for which CoMFA still yielded better results. Here again, the differences were present in both average results and optimum models. As the overall predictive ability of the models was degraded, the differences between the models were naturally also diminished. On average the models were also predictive as indicated by positive Pr-R^2 values, but the predictive ability of the models was strongly dependent on the compounds included in the training set as can be seen from the high standard deviation (SD) values of the Pr-R^2 indicators, whereas the Q^2 values are relatively stable suggesting that the QSAR model can usually be derived successfully based on the 2/3 of compounds. The largest change from the internal validation was observed for the *CALF* dataset, for which CoMFA gave slightly better average statistical indicators than BALL and even yielded the maximum Q^2 value. On the other hand, BALL generated the model with best external statistical indicators. All in all, the differences are certainly minor, but it is noteworthy that the relative predictive ability of BALL and CoMFA changed compared to the internal validation. The main reason for the poor average performance of BALL is the fact that it fails drastically for quite a few partitions of original data whereas it works very well for all other partitions. It seems that for some reason the *CALF* data tend to create labile models that, for some randomly selected partitions, lead to a reduced performance in the internal validation and to a total loss of external predictive ability.

For the SEAL superposition, the CoMFA results are uniformly rather poor, and the negative Pr-R^2 values indicate that the models actually have no predictive ability. Which, based on the poor internal performance, this is by no means surprising. As expected, the BALL results are considerably better, even though the SEAL superposed *CALF* dataset leads to relatively poor BALL models both in internal and in external validation.

When comparing the QSAR results (Tables 10, 11 and 12) one should bear in mind that CoMFA was used only as a benchmark and its parameters are not optimised, but even so BALL performs remarkably well considering that its descriptor vectors consist of few hundred elements rather than thousands, and it is thus much lighter technique than CoMFA. This implies that the BALL descriptors are faster to evaluate and much faster to process with statistical tools. This is primarily due to the design emphasis of the BALL technique, which is a grid-independent QSAR that could be easily automated for screening applications. The lack of grid, combined with the low-dimensional descriptors, may in part explain the slight preference BALL exhibits to the strictly constrained models as the part of the ligand that falls outside the template is not evaluated using a grid, but it is allocated to the template atoms in a fuzzy manner. If the ligand is much larger than the template or it has long protruding parts, the BALL descriptor will, by design, become fuzzy for that part and loose its accuracy when compared to the grid-based descriptors. Therefore the BALL benefits from the use of heavy weight factors which forcibly align the ligands on the template thus minimising the overspill.

3.5.3 Optimal BALL parameters

In general, the optimal areas of the BALL parameters 0.650-0.850 / 0.050-0.500 / 0.050-0.500 (min-max vdW_RI / min-max vdW_D / min-max EEL_D) found in this work are similar to the optimal areas 0.700-0.900 / 0.050-0.500 / 0.050-0.500 found in earlier validation work[465]. The VdW_D and EEL_D parameters have a higher variance, but overall they have a lesser impact on the Q^2 value. In particular, the EEL_D has only a slight effect on the predictive ability of the models as long as the compounds do not contain charged atoms. After a detailed analysis of the BALL models generated in this work, it became obvious that if any optimisation is to be performed on the vdW_RI parameter it should be restricted to the area of 0.650-0.850, as only seven optimal BALL models fall outside this range. Of those models, one belongs to *HUMANA*, one to *HUMANB* and five to *MOUSE*, and actually most of them could be easily replaced with comparable models belonging to the optimum area. For *HUMANA* dataset and FLUFF superposition **MIX CI EH**, the optimum BALL parameters are 550 / 500 / 250 (vdW_RI / vdW_D / EEL_D), but for the same superposition there is an alternative parameter set of 700 / 500 / 500, which produces only marginally lower Q^2 value (0.675 vs. 0.678). In the case of the *HUMANB* set the optimum parameters for **MIX CE IH** superposition with weight factors 1_1 are 25 / 250 / 500 and the nearest parameter set in the optimal range is 650 / 250 / 50 with a somewhat lower Q^2 value (0.368 vs. 0.401). On the other hand, if the vdW_RI values are restricted to the range 650-850, the optimal weight factors change to 3_2 and the optimal BALL parameters are 650 / 500 / 500, yielding a Q^2 value of 0.396.

For the troublesome *MOUSE* dataset a new set of optimal superpositions was generated by simply restricting the vdW_RI parameter to the range of 0.650-0.850. The statistical indicators of the QSAR models generated from this new set of superpositions are summarised in Table 13. The restriction of the vdW_RI caused many changes in statistical indicators, most notably the maximum Q^2 dropped from 0.611 to 0.519. Yet, the average Q^2 achieved suffered only a modest decrease from 0.497 to 0.475. So if the vdW_RI is restricted to the proposed range, no great loss of performance should ensue. Based on these findings it seems clear that a focused grid search in the area of vdW_RI 0.650-0.850 including the two dispersion parameters should yield nearly an optimum BALL model. On the other hand, the BALL parameter set of vdW_RI 0.800, vdW_D 0.500 and EEL_D 0.500, as already proposed in validation, should always provide a reasonable Q^2 value. For this diverse dataset BALL met or exceeded the results of the standard 3D-QSAR method CoMFA using either the tailor-made superposition technique FLUFF or the reference method SEAL. The FLUFF-BALL can easily be automated and as it is computationally simple, it provides a good computational “sieve” capable of fast screening of large molecular libraries.

Table 11. CALF, HUMANA and HUMANB internal validation results.

				Spress	Q^2	NPC	R^2	vdW_RI	vdW_D	EEL_D
CALF	**FLUFF**	**FIX**	**BALL**	0.718-0.788	0.431-0.433	8-15	0.826-0.966	700-750	50-75	50-250
	FLUFF	**FLEX**		0.622-0.718	0.518-0.695	8-15	0.841-0.980	650-800	50-250	50-250
	FLUFF	**MIX**		0.465-0.682	0.610-0.824	15	0.957-0.976	650-750	50-500	50-500
	FLUFF	**FIX**	**CoMFA**	0.730-0.915	0.405-0.502	11-15	0.967-0.999	-	-	-
		FLEX		0.680-0.683	0.502-0.530	10-12	0.976-0.980	-	-	-
		MIX		0.675-0.794	0.489-0.530	10-15	0.974-0.994	-	-	-
	SEAL	**-**	**BALL**	0.875	0.223	12	0.312	750	500	50
	SEAL	**-**	**CoMFA**	0.887	0.117	1	0.864	-	-	-
HUMANA	**FLUFF**	**FIX**	**BALL**	0.900-1.149	0.577-0.761	13-15	0.974-0.996	750-850	120-500	50-500
		FLEX		0.985-1.245	0.478-0.656	11-15	0.970-0.993	650-750	500	50-120
		MIX		1.024-1.209	0.569-0.678	13-15	0.984-0.999	550-750	50-500	120-250
	FLUFF	**FIX**	**CoMFA**	1.131-1.481	0.353-0.407	2-15	0.609-1.000	-	-	-
		FLEX		1.148-1.216	0.294-0.400	3-6	0.728-0.935	-	-	-
		MIX		1.189-1.231	0.289-0.333	2-5	0.513-0.915	-	-	-
	SEAL	**-**	**BALL**	1.378	0.375	15	0.998	650	120	500
	SEAL	-	**CoMFA**	1.384	0.163	2	0.461	-	-	-
HUMANB	**FLUFF**	**FIX**	**BALL**	1.019-1.049	0.398-0.606	4-15	0.658-0.991	650-750	500	50-500
		FLEX		1.026-1.345	0.382-0.486	10-15	0.954-0.997	650-750	500	50-500
		MIX		1.137-1.213	0.397-0.533	13-15	0.971-0.993	25-750	250-500	120-500
	FLUFF	**FIX**	**CoMFA**	1.102-1.392	0.313-0.338	3-15	0.706-1.000	-	-	-
		FLEX		1.078-1.445	0.286-0.376	3-15	0.770-1.000	-	-	-
		MIX		1.072-1.147	0.281-0.383	2-5	0.529-0.901	-	-	-
	SEAL	**-**	**BALL**	1.039	0.410	4	0.801	700	120	50
	SEAL	-	**CoMFA**	1.272	0.279	8	0.965	-	-	-

Table 12. MOUSE and RAT internal validation results.

				Spress	Q^2	NPC	R^2	vdW_RI	vdW_D	EEL_D
MOUSE	**FLUFF**	**FIX**	**BALL**	1.154-1.270	0.476-0.611	9-15	0.871-0.905	250-550	250-500	75-250
	FLUFF	**FLEX**		1.195-1.282	0.438-0.512	6-7	0.771-0.808	850-900	50-500	250-500
	FLUFF	**MIX**		1.197-1.251	0.431-0.519	2-7	0.543-0.803	700-850	50-250	50-120
	FLUFF	**FIX**	**CoMFA**	1.232-1.303	0.410-0.482	5-6	0.899-0.941	-	-	-
		FLEX		1.312-1.362	0.325-0.374	2-4	0.544-0.883	-	-	-
		MIX		1.337-1.344	0.353-0.427	3-10	0.702-0.986	-	-	-
	SEAL	**-**	**BALL**	1.345	0.362	4	0.635	500	50	75
	SEAL	**-**	**CoMFA**	1.592	0.178	2	0.464	-	-	-
RAT	**FLUFF**	**FIX**	**BALL**	1.257-1.334	0.470-0.533	7-8	0.719-0.776	800-850	250-500	50-250
		FLEX		1.262-1.396	0.452-0.529	7-15	0.736-0.914	700-850	120-500	120-500
		MIX		1.243-1.367	0.490-0.547	7-15	0.690-0.962	800-850	120-500	75-500
	FLUFF	**FIX**	**CoMFA**	1.060-1.174	0.582-0.673	5-10	0.867-0.963	-	-	-
		FLEX		1.245-1.337	0.445-0.545	5-11	0.846-0.967	-	-	-
		MIX		1.226-1.332	0.463-0.546	4-12	0.824-0.977	-	-	-
	SEAL	**-**	**BALL**	1.419	0.385	4	0.583	600	50	250
	SEAL	-	**CoMFA**	1.743	0.157	2	0.351	-	-	-

Table 13. MOUSE internal validation results when vdW_RI values are restricted to the range of 0.650-0.850.

			Spress	Q^2	NPC	R^2	F	vdW_RI	vdW_D	EEL_D
FLUFF	**BALL**	**FIX**	1.203-1.275	0.459-0.482	3-10	0.605-0.965	34.9-45.6	650-800	50-120	75-120
		FLEX	1.195-1.282	0.438-0.512	6-7	0.756-0.819	34.8-37.9	850	50-500	250-500
		MIX	1.197-1.251	0.431-0.519	2-7	0.768-1.000	35.2-39.2	700-850	50-250	50-120

4. MCSOR: A PLS-TYPE HYBRID ALGORITHM

A promising 3D QSAR method, Self-Organizing Molecular Field Analysis (SOMFA), has recently been introduced[200] and applied to medicinal chemistry[483-488] and even to food chemistry[489]. SOMFA has not become a widely adopted QSAR technique despite of its conceptual simplicity and easy implementation. The basis of the SOMFA technique is similar to the CoMFA[14] and GRID[15] in the sense that it also uses a grid of points to generate the descriptor. However, SOMFA does not use probe interaction energies like CoMFA does but instead it relies upon descriptors directly derived from the intrinsic molecular properties, such as shape and electrostatic potential calculated from partial charges. One of the novel features of the SOMFA algorithm is the fact that unlike most of the other QSAR techniques SOMFA comes with a built-in regression methodology instead of relying on an external method such as PLS.

In SOMFA regression each of the 3D descriptor grids (*D*) are multiplied by the mean centred observed activity (eq. 60) which is derived by subtracting the mean of the training set from each value *Y*, whereupon the largest dependent variables will have positive values and the smallest will have negative values. These individually multiplied grids are summed to for the so-called master matrix (*MM*) which contains the relative weights of each descriptor variable.

$$MM = \sum_{i=1}^{i=n} D_i \left(Y_i - \bar{Y} \right) \tag{60}$$

In the next phase the master matrix is used to reduce the original descriptor grid into a single number (eq. 61). These numbers are in turn used to derive a MLR model of the correlation between structure and activity. The SOMFA regression approach is computationally very simple and it can be used to create visualisations outlining the relative importance of different molecular features[200].

$$d_i = D_i MM' \tag{61}$$

During our work with a SOMFA application[485] 3D SOMFA grids were transformed into vectors in order to make large cross-validation runs more efficient. This led to the interesting observation that the basic principle of SOMFA regression tool, which could be called *Self-Organising Regression* (SOR), applies to many underdetermined regression problems (i.e. ones in which the number of variables is much larger than the number of objects) frequently encountered in different areas of QSAR.

4.1 From SOR to MCSOR

Mathematically, the basic principle of SOR is the same as that of SOMFA, i.e., crucial to SOR is the concept of mean-centred data. In original SOMFA methodology the descriptors for each object in the training set are combined to form the master grid. For SOR this grid is replaced by a *mastervector* (*MV*, eq. 62) which is created by summing the descriptors, element by element, after scaling each descriptor (row of *X*) by the corresponding value of mean centred dependent variable (Y_0). The independent variable block *X* can also be mean centred to form X_0.

$$MV = \sum_{i=1}^{i=n} X(i) \cdot Y_0(i) \tag{62}$$

where the subscript *i* refers to the sample (row). Thus the dimensionality of the master vector is exactly the same as that of the descriptor. A predictive equation relating the descriptors (independent variables) to the values of *y* (dependent variable) can be derived from the master vector in three steps. First, for every object in the training set, a predictor (regression variable, *P*) is calculated using eq. 63

$$P(i) = X(i) \cdot MV' \tag{63}$$

i.e., as a dot product of the descriptor and master vector (the apostrophe stands for a vector transpose). Second, a univariate regression model (MLR) is derived using the predictor vector as an independent variable. Third, eq. 63 and the regression coefficients (*B*) derived in the second phase are used to calculate an estimate for the dependent variable ($\hat{Y}$) for each object in the test set. In this way, the information contained in the high-dimensional descriptors can be compressed into a single variable. Yet, in its current form the SOR algorithm and the original SOMFA implementation are limited to univariate-Y problems.

In order to ensure that the internal and external predictions are truly 'blind', however, the master vector must be calculated for each training set separately (i.e. the master vector should not be 'contaminated' by the descriptors or activities of the test set molecules), after which the corresponding regression model can be derived. Note that this remark also applies to Leave-One-Out cross-validation – otherwise internal predictions would seem to work even with random numbers.

However novel the SOR regression method may seem to be, it can be demonstrated that the SOR is mathematically identical to the single principal component implementation of SIMPLS[350] partial least-squares approach. To elaborate this point a detailed description of the SOR algorithm is given in Chart 1 and a similar description of SIMPLS can be found in Chart 2. When one compares the SOR and SIMPLS algorithms it becomes evident that if the descriptors are mean centred (line 2), then $MV = Y_0^{'} X_0$ and $S = X_0^{'} Y_0$. Furthermore, as $A'B = \left(AB' \right)'$ it also holds that $Y_0^{'} X_0 = (X_0^{'} Y_0)'$ and therefore *MV=S'* (line 3). Also due to this equivalence the *P=t* (lines 4-6).

Chart 1. Detailed SOR algorithm.

1 $Y_0 = Y - \overline{Y}$

2 $X_0 = X - \overline{X}$

3 $MV = \sum_{i=1}^{i=n} X_0(i) Y_0(i) = Y_0^{'} X_0$

4 $P = X_0 MV'$

5 $P = P - \overline{P}$

6 $P = \frac{P}{\|P\|}$

7 $B_{SOR} = \left(P'P\right)^{-1} P'Y_0$

8 $\hat{Y}_{SOR} = PB_{SOR} + \overline{Y}$

Chart 2. Detailed SIMPLS algorithm.

1 $Y_0 = Y - \overline{Y}$

2 $X_0 = X - \overline{X}$

3 $S = X_0^{'} Y_0$

4 $t = X_0 S$

5 $t = t - \overline{t}$

6 $t = \frac{t}{\|t\|}$

7 $q = Y_0^{'} t$

8 $B_{SIMPLS} = \left(\frac{S}{\|X_0 S\|}\right) q'$

9 $\hat{Y}_{SIMPLS} = X_0 B_{SIMPLS} + \overline{Y}$

If one then compares the B_{SOR} with the q (line 7), while keeping in mind that $||P||=1$, it follows that $(P'*P)^{-1}=1$ and hence

$$B_{SOR} = \left(P'P\right)^{-1} P'Y_0 = P'Y_0 \tag{64}$$

as $q = Y_0^{'} t$ and $P=t$ it follows that

$$B_{SOR} = P'Y_0 = (Y_0^{'} P)' = (Y_0^{'} t)' = q' \tag{65}$$

By applying the previous result to $\hat{Y}_{SOR}$ (line 8), and as P=t one can utilise the definition of t (lines 4-6) from the SIMPLS algorithm, it directly follows that

$$\hat{Y}_{SOR} = PB_{SOR} + \overline{Y} = tq' + \overline{Y} = \left(\frac{(X_0 S)}{\|X_0 S\|}\right) q' + \overline{Y} \tag{66}$$

In the SIMPLS algorithm one can combine the definition of the $\hat{Y}_{SIMPLS}$ (line 9) with the definition of B_{SIMPLS} (line 8) to create eq. 67.

$$\hat{Y}_{SIMPLS} = X_0 \left(\frac{S}{\|X_0 S\|}\right) q' + \overline{Y} \tag{67}$$

When the equation for $\hat{Y}_{SOR}$ (eq. 66) and $\hat{Y}_{SIMPLS}$ (eq. 67) are compared, it is evident that

$$\hat{Y}_{SOR} = \left(\frac{(X_0 S)}{\|X_0 S\|} \right) q' + \bar{Y} = X_0 \left(\frac{S}{\|X_0 S\|} \right) q' + \bar{Y} = \hat{Y}_{SIMPLS} \tag{68}$$

which proves beyond any doubt that the SOR model is, in fact, equivalent to a single component SIMPLS model.

Therefore it stands to reason that more complex datasets requiring multiple components to achieve sufficient predictive power will fail if they are analysed using the SOR methodology. Consequently, it is necessary to extend the original SOR methodology to multiple components. In other words, one must formulate a *MultiComponent Self-Organizing Regression* (MCSOR) if the SOR principle is to be used for more complex datasets. The MCSOR uses the errors, also called residuals, of the preceding component as dependent variable and derives additional mastervectors which can then be used in conjunction with MLR to generate a multicomponent model. A detailed pseudocode representation of the MCSOR algorithm is presented in Chart 3. For the first component the mastervector is computed and the independent block is reduced to single value just as in normal SOR (lines 1-2) and then MLR is used to derive the regression model (B) of dependent variable (Y_0) including the intercept value (lines 3-4). The mastervector (MV), regression coefficients (B) and the estimate of dependent variable ($\hat{Y}$) are added to mastermatrix (MM), betamatrix (BM) and Y_{pred}, respectively (lines 5-7).

The additional components are then derived by replacing the original dependent variable Y_0 with the error of prediction (line 9). This new Y is then used to derive a new mastervector which is used to create a new set of X_{pred} values (lines 9-11). This new X-vector and the preceding X-vector(s) are processed with MLR to produce a new regression model of Y (line 12). The mastervector (MV) and regression coefficients (B) are added to mastermatrix (MM) and betamatrix (BM), respectively (lines 5-7). The new regression model is used to compute new predicted values (lines 15-18). Finally the estimate of dependent variable ($\hat{Y}$) is added to Y_{pred}. This process is repeated each time replacing the earlier error of prediction with the new values until a desired number of components (NPC) have been extracted (lines 8-20 are iterated). Even though the MCSOR can derive multiple SOR components it is still limited to univariate-Y problems because it directly depends on the SOR algorithm.

Chart 3. Detailed MCSOR algorithm.

1 $MV = Y_0^{'} * X_0$

2 $X_{pred} = X_0 * MV'$

3 $B = MLR(X_{pred}, Y_0)$

4 $\hat{Y} = B(2) * X_{pred} + B(1)$

5 $BM(1,1:2) = B$

6 $MM(1) = MV$

7 $Y_{pred}(1) = \hat{Y}$

8 *for* $j = 2 : NPC$

9 $Y = Y_0 - \hat{Y}$

10 $MV = Y' * X_0$

11 $X_{pred} = X_0 * MV'$

12 $B = MLR(X_{pred}, Y)$

13 $MM(j) = MV$

14 $BM(j,1:j+1) = B$

15 $\hat{Y} = B(1)$

16 *for* $k = 1 : j$

17 $\hat{Y} = \hat{Y} + B(k+1) * X_0 * MM(k)'$

18 *end*

19 $Y_{pred}(j) = \hat{Y}$

20 *end*

4.2 SOMFA using MCSOR and other multivariate methods

The *CBG*, *MCF log* K_a, *MCF* pEC_{50}, *PCDD* and *PCDF* sets, already familiar from the validation of FLUFF-BALL, are used to evaluate the effect of the different statistical techniques on the performance of the SOMFA. For the other sets the same molecular models, including templates, (for details see page 71) were used for two rigid superpositions generated using the SEAL and FLUFF **FIX** methodologies. Semi-rigid and flexible superpositions were also performed using the FLUFF **MIX** and **FLEX** methodologies. After the superposition the SOMFA steric and electrostatic descriptors were computed as described in the original article[200] using a 22Å cubic grid with the granularity of 0.5Å. The evaluation of SOMFA descriptors and statistical computations were performed using MATLAB[461] scripts.

In order to reliably estimate the performance difference between the statistical methods it is essential to minimise the effect of chance. Therefore, 2500 randomly generated divisions to training and testing sets were generated for each of the datasets. From *CBG* data 21 molecules were selected for the training set and 10 for test set and for both *MCF* sets the similar division was 28 and 14 molecules, respectively. In case of PCDD and *PCDF* data 7 compounds were separated for test set leaving 19 and 27 compounds respectively for the training set. SOR, MCSOR and SIMPLS were then used to build regression models while the maximum number of components for MCSOR and SIMPLS models was set to 7 for both of the *MCF* sets and for *PCDF* and to 5 for the CBG and *PCDD* sets.

When the average statistical descriptors of SOR models derived from the 5 datasets (Table 14) are compared with the corresponding values of MCSOR or SIMPLS (Table 15 and Table 16) it becomes evident that in general MCSOR and SIMPLS are clearly superior to SOR. Even in the case of the relative simple *PCDD* and *PCDF* datasets the SOR generates very poor models whereas the MCSOR and SIMPLS are able to generate clearly predictive models as indicated by the high Pr-R^2 values. The performances of MCSOR and SIMPLS are virtually identical for all datasets. The only dataset for which the SOR generates a model comparable to MCSOR and SIMPLS is the *CBG* set which has become the *de facto* benchmark dataset with which new QSAR techniques are tested. The good performance of SOR indicates that a predictive model can be created for this dataset using only one principal component and additional components contribute only a very minor increase in predictive power. This is in line with the criticism of several authors[490-492] who point out that the *CBG* is overly simple set as it can be explained by single component regression model and more alarmingly almost all QSAR techniques are able to derive highly predictive models from it. Thus, the CBG data set represents an ideal case in that there is no structural Y-variation in the column space of X. Unfortunately a vast majority of QSAR problems do contain structured noise and therefore the CBG data set is not particularly suitable for benchmarking, at least to test feature selection methods. However, it has served repeatedly as a preliminary test set for QSAR methods[14,20,24,200,202,220,266,269,270,272,279,289,493-495]. In the case of SOR, the *CBG* was used as a benchmark dataset in the original SOMFA paper, and as this particular dataset works very well with only one principal component, the results gave an unrealistically good picture of SOR's performance.

Table 14. Average SOR statistical descriptors over the 2500 random divisions of CBG, PCDD, PCDF, MCF pEC_{50} and MCF log K_a datasets.

			S_{press}	Q^2	R^2	R^2_{ex}	SDEP	Pr-R^2
CBG	**FLUFF**	**FIX**	0.728	0.560	0.683	0.640	0.683	0.601
	FLUFF	**FLEX**	0.686	0.608	0.730	0.703	0.625	0.660
	FLUFF	**MIX**	0.665	0.633	0.746	0.684	0.635	0.645
	SEAL		0.727	0.563	0.685	0.625	0.688	0.584
PCDD	**FLUFF**	**FIX**	1.574	-0.128	0.421	0.405	1.418	-0.139
	FLUFF	**FLEX**	1.568	-0.109	0.468	0.436	1.344	-0.150
	FLUFF	**MIX**	1.550	-0.100	0.468	0.407	1.398	-0.065
	SEAL		1.564	-0.130	0.415	0.368	1.470	-0.159
PCDF	**FLUFF**	**FIX**	1.271	0.193	0.562	0.444	1.154	0.172
	FLUFF	**FLEX**	1.186	0.289	0.673	0.476	1.074	0.315
	FLUFF	**MIX**	1.173	0.306	0.681	0.452	1.102	0.282
	SEAL		1.184	0.290	0.680	0.452	1.090	0.286
MCF pEC_{50}	**FLUFF**	**FIX**	1.432	-0.071	0.285	0.154	1.366	0.012
	FLUFF	**FLEX**	1.433	-0.051	0.285	0.175	1.316	0.052
	FLUFF	**MIX**	1.445	-0.081	0.276	0.156	1.346	0.024
	SEAL		1.434	-0.081	0.280	0.141	1.363	0.030
MCF log K_a	**FLUFF**	**FIX**	1.143	0.030	0.347	0.191	1.078	0.116
	FLUFF	**FLEX**	1.128	0.055	0.347	0.211	1.072	0.135
	FLUFF	**MIX**	1.120	0.063	0.352	0.195	1.079	0.128
	SEAL		1.134	0.040	0.352	0.192	1.090	0.109

Table 15. Average MCSOR statistical descriptors over the 2500 random divisions of CBG, PCDD, PCDF, MCF pEC_{50} and MCF log K_a datasets.

			S_{press}	Q^2	NPC	R^2	R^2_{ex}	SDEP	Pr-R^2
CBG	**FLUFF**	**FIX**	0.689	0.641	2.9	0.844	0.651	0.690	0.581
	FLUFF	**FLEX**	0.673	0.631	1.6	0.773	0.684	0.664	0.590
	FLUFF	**MIX**	0.653	0.655	1.7	0.788	0.648	0.686	0.566
	SEAL		0.681	0.648	2.8	0.846	0.639	0.685	0.578
PCDD	**FLUFF**	**FIX**	0.932	0.665	3.9	0.859	0.741	0.823	0.568
	FLUFF	**FLEX**	0.874	0.706	3.8	0.891	0.756	0.746	0.582
	FLUFF	**MIX**	0.868	0.706	3.9	0.890	0.772	0.745	0.639
	SEAL		0.940	0.656	4.0	0.856	0.732	0.876	0.573
PCDF	**FLUFF**	**FIX**	0.953	0.621	5.7	0.880	0.670	0.819	0.448
	FLUFF	**FLEX**	0.804	0.713	4.2	0.844	0.710	0.752	0.621
	FLUFF	**MIX**	0.802	0.714	4.1	0.842	0.692	0.771	0.601
	SEAL		0.809	0.710	4.3	0.846	0.683	0.761	0.602
MCF pEC_{50}	**FLUFF**	**FIX**	1.185	0.318	4.9	0.757	0.268	1.299	0.260
	FLUFF	**FLEX**	1.176	0.306	4.7	0.727	0.332	1.189	0.307
	FLUFF	**MIX**	1.216	0.312	5.3	0.778	0.281	1.238	0.262
	SEAL		1.203	0.303	5.2	0.783	0.271	1.281	0.248
MCF log K_a	**FLUFF**	**FIX**	0.971	0.408	4.8	0.797	0.475	0.845	0.460
	FLUFF	**FLEX**	0.970	0.387	4.5	0.763	0.451	0.863	0.441
	FLUFF	**MIX**	0.953	0.432	5.1	0.807	0.442	0.885	0.424
	SEAL		0.948	0.426	4.8	0.807	0.452	0.854	0.457

Table 16. Average SIMPLS statistical descriptors over the 2500 random divisions of CBG, PCDD, PCDF, MCF pEC_{50} and MCF log K_a datasets.

			S_{press}	Q^2	NPC	R^2	R^2_{ex}	SDEP	Pr-R^2
CBG	**FLUFF**	**FIX**	0.700	0.631	3.0	0.848	0.653	0.691	0.583
	FLUFF	**FLEX**	0.688	0.615	1.6	0.772	0.680	0.668	0.599
	FLUFF	**MIX**	0.667	0.642	1.8	0.790	0.646	0.689	0.567
	SEAL		0.695	0.635	2.9	0.849	0.642	0.681	0.590
PCDD	**FLUFF**	**FIX**	0.939	0.664	4.1	0.866	0.742	0.814	0.587
	FLUFF	**FLEX**	0.884	0.701	3.9	0.892	0.751	0.752	0.592
	FLUFF	**MIX**	0.878	0.700	3.9	0.892	0.767	0.750	0.646
	SEAL		0.946	0.655	4.1	0.862	0.733	0.872	0.582
PCDF	**FLUFF**	**FIX**	0.945	0.630	5.8	0.883	0.673	0.810	0.472
	FLUFF	**FLEX**	0.808	0.711	4.3	0.845	0.709	0.754	0.620
	FLUFF	**MIX**	0.807	0.711	4.2	0.844	0.687	0.770	0.603
	SEAL		0.812	0.709	4.4	0.848	0.683	0.762	0.607
MCF pEC_{50}	**FLUFF**	**FIX**	1.182	0.318	5.0	0.764	0.275	1.314	0.244
	FLUFF	**FLEX**	1.182	0.306	4.8	0.734	0.338	1.178	0.297
	FLUFF	**MIX**	1.218	0.308	5.3	0.777	0.294	1.261	0.251
	SEAL		1.206	0.300	5.3	0.785	0.289	1.311	0.243
MCF log K_a	**FLUFF**	**FIX**	0.979	0.412	5.2	0.816	0.479	0.852	0.454
	FLUFF	**FLEX**	0.971	0.390	4.7	0.773	0.459	0.867	0.428
	FLUFF	**MIX**	0.961	0.433	5.5	0.826	0.443	0.898	0.406
	SEAL		0.958	0.429	5.2	0.816	0.463	0.857	0.443

4.3 MultiComponent SOMFA on xenoestrogen datasets

In order to further evaluate the performance difference between SOR and other multivariate methods the SOMFA descriptors of Cramer Testosterone data (*TBG*) and Sadler (*SADLER*) sets were used to build QSAR models with SOR, MCSOR and SIMPLS. The TBG set, introduced by Cramer in the original CoMFA paper[14], contains 21 molecules whose binding affinities to testosterone binding globulin are known, whereas the *SADLER* dataset[63,496] contains 30 compounds with a considerable binding affinity to the oestrogen receptor. Also, the same xenoestrogen dataset that was used to test the FLUFF-BALL methodology was also analysed using the SOMFA QSAR (for details see page 78). Furthermore, the performance of a less known SOMFA descriptor, proposed by Bradley and Waller[497], based on a molecular polarisability field was evaluated as a stand-alone descriptor but also in conjunction with the standard steric and electrostatic descriptors.

After superposition there were 12 FLUFF alignments (**FIX**, **MIX** and **FLEX** with **CI**/**CE** and **IH**/**EH** modifications) and a reference SEAL superposition for the *TBG* and *SADLER* datasets. On the other hand the introduction of weight factors in the EDKB datasets resulted in a total of 132 different FLUFF superposed sets and the reference SEAL. All sets were centred and a 22Å cubic grid with granularity of 0.5Å was created. SOMFA descriptors were then evaluated using MATLAB[461] scripts. The SOMFA steric (*SHAPE*, eq. 69) and electrostatic (*ESP*, eq. 70) descriptors were computed as described in the original SOMFA article[200]. A field describing the polarisability of the molecule[497] (*POLAR*, eq. 71) was also evaluated. The atomic polarisabilities required were obtained using the method of Lewis[498] and an in-house modified version of AMPAC program (version 2.1, QCPE#506).

$$SHAPE(p) = \sum_n \begin{cases} 1 & r_{p,a_n} < r_{vdw} \\ 0 & r_{p,a_n} \geq r_{vdw} \end{cases} \tag{69}$$

$$ESP(p) = \sum_n \frac{Q_{a_n}}{r_{p,a_n}^2} \tag{70}$$

$$POLAR(p) = \sum_n \frac{P_{a_n}}{r_{p,a_n}^3} \tag{71}$$

where $\boldsymbol{p}$ is a grid point and $\boldsymbol{a}$ is an atom and $\boldsymbol{r_{a,p}}$ is the distance between p and a, $\boldsymbol{r_{vdw}}$ is the van der Waals radius of atom a, $\boldsymbol{Q_a}$ is the partial charge of a and the $\boldsymbol{P_a}$ is the atomic polarisability of atom a.

SOR[200,499], MCSOR[499] and SIMPLS[350] regression models were generated for SHAPE, ESP and POLAR descriptors and their combinations (SHAPE_ESP, SHAPE_POLAR, ESP_POLAR and SHAPE_ESP_POLAR) using Leave-one-out cross-validation (LOO CV) and in-house MATLAB[461] scripts. The maximum number of principal components was set at 7 for *TBG* and

SADLER and at 15 for *EDKB*. For the *RAT* set the maximum number of principal components could be considerably higher and therefore additional MCSOR and SIMPLS models with the maximum number of principal components set at 32 were generated. Some increase in the Q^2 values was observed, but the benefits were negligible (<0.020) and at the same time the number of components rose dramatically, being in the range from 27 to 31. Therefore it was judged that the gains made in the Q^2 values did not outweigh the dramatic increase in the number of principal components and so these models were discarded.

When the results superposition results were analysed it soon became evident that the SOMFA models are gratifyingly stable in regard to the superposition as the FLUFF weight factors as CE/CI and EH/IH modifications had only a minor impact on the Q^2 values (typically <5%). Also, the FIX, FLEX and MIX variants yielded similar results (typical difference <10%) so that in the summarised internal validation results (Tables 19- 23) only the range of statistical descriptors yielded by the FLUFF models is shown. Due to the neglible effect the superposition methods had on the overall performance of the models, no clear trend pointing to an optimum superposition could be ascertained from the data. In general the effect of the FLUFF weight factor was less than ~10% of the average Q^2 value. The only notable exception to this was the *CALF* dataset where the variation was much higher but this could be explained by the rather poor performance of this dataset whereby a small variation of the Q^2 (±0.050) could lead to a very high relative variation (~30-50%). The effect of the CE/CI and EH/IH modifications was even weaker averaging to less than 5% and once again the CALF dataset yielded higher variation. The FLUFF FIX, FLEX and MIX variants also created very similar models (Q^2 difference ~10%) but this time there was not any clear difference between the *CALF* and other datasets. Also the performance of SEAL superposition was similar to that of the FLUFF and therefore no clear distinction between performances of the two techniques could be made.

When the relative performance of different descriptors is analysed no clear pattern emerges. In the case of the benchmark datasets the *TBG*, which is known to be computationally difficult, the SHAPE and POLAR descriptors work quite well whereas the ESP fails. From the results it is clear that for this dataset no major increase in predictivity is obtained by combining the descriptors. The Q^2 values (0.133-0.563) of the TBG set are comparable to those reported for CoMFA[14,458] (0.555 and 0.601) whereas the results for region focused CoMFA[500] and COMPASS[279] are superior (0.658 and 0.88, respectively). In the case of COMSA[202] the range of Q^2 reported is large (0.15-0.76) and for the most part the SOMFA results are comparable but there are some COMSA models which yield clearly superior Q^2 values. For the SADLER set all of the descriptors yield rather good models, even though it is evident that the ESP descriptor works particularly well. However, the combined models ESP POLAR and SHAPE ESP POLAR yielded best models, which suggest that the information contained by the ESP descriptor can be augmented by the POLAR descriptor. The Q^2 values of the *SADLER* dataset (0.409-0.698) are comparable to the values yielded by CoMFA[496](0.537-0.720) and better than the values reported for receptor interaction energy based QSARs[63] (0.487-0.570). On the other hand, the results are somewhat lower than the ones reported for CoMFA optimised with a region focusing technique[496] (0.651-0.796) and significantly lower than the ones yielded by GRID models utilising receptor based alignment and region focusing[63] (0.830 and 0.921).

When results of the different EDKB datasets and SOMFA descriptors are summarised the *CALF* set emerges as the most troublesome of the five EDKB sets, which is clearly indicated by the uniformly low Q^2 values (0.052-0.290), especially if one takes into account the Q^2 values of (0.54 and 0.61) reported in the literature[479]. For this set the optimal descriptors seem to be SHAPE and ESP whereas the combined descriptors SHAPE_ESP and SHAPE_ESP_POLAR generate somewhat lower values. For *HUMANA*, *HUMANB* the POLAR field yielded best results. SHAPE and ESP also generate valid models. All of the combined models also work but the performance is lower than that of the single descriptors. Especially the SHAPE_ESP combination seems to lead to considerable loss of predictive power. In the *MOUSE* dataset there is a slight difference between the results of the MCSOR and SIMPLS regression methods. Both regression techniques are able to derive highly predictive models from all descriptors but the MCSOR yields slight better model for ESP descriptor and the difference is considerable for POLAR descriptor (0.431-0.514 vs. 0.342-0.371). On the other hand the difference is lost in the combined descriptors. The Q^2 values of the *MOUSE* (0.282-0.554) are slightly lower, but still comparable to the CoMFA and HQSAR results reported in the literature[122] (0.59 and 0.58, respectively) but significantly lower than FRED/SKEYS (0.70) or kNN (0.77) results [122,501]. The POLAR and ESP fields along with the combined ESP_POLAR and SHAPE_ESP_POLAR fields yield the best models for *RAT* set but reasonable models could be derived from all SOMFA fields. Yet, the results are inferior to a CoMFA results reported in the literature[502] as indicated by the Q^2 values of 0.334-0.541 yielded by SOMFA and the value 0.71 yielded by CoMFA. In general, the differences between the performances of the descriptors are very small, and are based upon a limited set of molecules, so no universal recommendations can be made at this time. On the other hand, the results indicate that the polarisability descriptor as proposed by Bradley et al[497], while relatively easy to compute, manages to produce valid and predictive descriptors for a diverse xenoestrogen dataset. Therefore it should be considered along the steric and electrostatic descriptors to be one of the standard descriptors used a SOMFA QSAR analysis.

For each of the EDKB datasets Y-scrambling and bootstrapping runs were performed using SOR, MCSOR and SIMPLS as described in section 3.5.2 (see page 84). For *TBG* and *SADLER* datasets the division to test and training set was 5/21 and 12/36, respectively. In all Y-scrambling cases the predictive ability was completely lost while in the bootstrapping runs the statistical performance indicators worsened, but all models still produced reasonable Q^2 values. As the overall predictive ability of the models was degraded, the differences between the models were naturally also diminished. Nevertheless, on average the models were still predictive as indicated by positive Pr-R^2 values and the relative differences in predictive ability were also preserved.

Even in the external validation the performance of *CALF* was very low and it also seems that for some partitions a labile model is created which leads to a reduced performance in the internal validation and to a total loss of external predictive ability. This behaviour, when combined with the poor results of the internal validation merit a further analysis. First of all the possibility of artefacts arising from the placement of molecules in the SOMFA grid was analysed by creating a set of new descriptors where the molecules were rotated along the x and y axis at 5^o intervals up to 90^o. Also the effect of 0.25Å translation along the x, y and z axis were performed.

None of these modifications yielded any significant improvement in the Q^2 values. Further tests included an extended SOMFA grid with the same 0.5Å granularity, but with the span of 44Å, and a denser grid with the granularity of 0.25Å, but none of these new models yielded any major improvement to the low Q^2 values. The LOO results of the *CALF* dataset were then analysed to see whether the poor Q^2 is caused by very poor prediction of few compounds, but that was not the case. However, the observed values of *CALF* data are, for the most part, clustered around the mean value of 0.40 and there are only relatively few values outside of the central cluster. Therefore, a set of models was generated where some of these points were excluded to see if this would yield higher Q^2 values, but the exclusion of outlying points did not increase the predictivity of the model. As a result, the reason for the poor performance of *CALF* dataset unfortunately remains a mystery.

When comparing different regression tools it is quite obvious that the SOR creates inferior models when compared to MCSOR and SIMPLS. Yet, in case of the *MOUSE* dataset the difference is less marked than in the other four sets. All in all these results clearly indicate that for a diverse dataset SOMFA clearly benefits from the use of external regression tools instead of the SOR regression, which is actually a common PLS with only one principal component[200]. The performance difference between SOR and external regression techniques is most likely due by the well-known fact that more than one principal component is required to accurately describe most QSAR datasets[5,458,503-507]. The overall performance of MCSOR and SIMPLS is very similar, though for some reason the ESP field of HUMANA and HUMANB seems to favour MCSOR over SIMPLS as indicated by Q^2 ranges 0.445-0.515 vs. 0.208-0.267 and 0.366-0.417 vs. 0.065-0.152, for *HUMANA* and *HUMANB*, respectively. A similar effect can also be observed in the *MOUSE* POLAR descriptor for which the Q^2 ranges are 0.431-0.514 vs. 0.342-0.371, for MCSOR vs. SIMPLS, respectively. In the bootstrapping runs the SOR still generates inferior models and the unexpected performance difference observed between MCSOR and SIMPLS is preserved for the ESP field of *HUMANA* and *HUMANB* datasets while the difference is lost in the case of *MOUSE* POLAR field. Therefore it is likely that the difference in performance is highly dependent on the exact composition of the dataset and this difference may not be significant for general case. As the performance of MCSOR and SIMPLS is almost identical and the SIMPLS is computationally lighter than MCSOR it would advocate the use SIMPLS as external regression tool for SOMFA analysis.

Table 17. *TBG* internal validation results. For FLUFF the range (min – max) of values generated by different superpositions is given.

			S_{press}	Q^2	NPC	R^2
SOR	**SHAPE**	**FLUFF**	1.145-1.190	0.080-0.148	-	0.441-0.461
		SEAL	1.143	0.151	-	0.500
	ESP	**FLUFF**	1.399-1.436	-0.341- -0.272	-	0.244-0.250
		SEAL	1.419	-0.309	-	0.276
	POLAR	**FLUFF**	1.019-1.025	0.317-0.326	-	0.448-0.454
		SEAL	1.053	0.279	-	0.431
	SHAPE ESP	**FLUFF**	1.404-1.411	-0.294- -0.281	-	0.259-0.272
		SEAL	1.397	-0.268	-	0.288
	SHAPE POLAR	**FLUFF**	1.025-1.025	0.317-0.318	-	0.448-0.450
		SEAL	1.054	0.277	-	0.429
	ESP POLAR	**FLUFF**	1.052-1.060	0.269-0.281	-	0.454-0.460
		SEAL	1.088	0.231	-	0.440
	SHAPE ESP POLAR	**FLUFF**	1.052-1.060	0.270-0.281	-	0.456-0.462
		SEAL	1.087	0.231	-	0.443
MCSOR	**SHAPE**	**FLUFF**	0.988-1.111	0.367-0.499	5-5	0.966-0.984
		SEAL	0.971	0.516	5	0.971
	ESP	**FLUFF**	1.259-1.280	0.133-0.184	4-5	0.711-0.863
		SEAL	1.165	0.257	4	0.757
	POLAR	**FLUFF**	0.924-1.031	0.382-0.533	3-4	0.768-0.861
		SEAL	1.060	0.424	5	0.904
	SHAPE ESP	**FLUFF**	1.188-1.241	0.210-0.234	4-5	0.769-0.897
		SEAL	1.129	0.303	4	0.811
	SHAPE POLAR	**FLUFF**	0.893-1.020	0.395-0.563	3-4	0.787-0.875
		SEAL	1.038	0.447	5	0.925
	ESP POLAR	**FLUFF**	0.896-0.940	0.486-0.533	3-3	0.828-0.829
		SEAL	0.957	0.467	3	0.834
	SHAPE ESP POLAR	**FLUFF**	0.892-0.935	0.491-0.537	3-3	0.833-0.835
		SEAL	0.951	0.474	3	0.840
SIMPLS	**SHAPE**	**FLUFF**	1.048-1.194	0.268-0.436	5-5	0.966-0.984
		SEAL	1.034	0.451	5	0.971
	ESP	**FLUFF**	1.257-1.274	0.121-0.190	4-5	0.711-0.863
		SEAL	1.174	0.246	4	0.757
	POLAR	**FLUFF**	1.015-1.137	0.248-0.436	3-4	0.768-0.861
		SEAL	1.116	0.361	5	0.904
	SHAPE ESP	**FLUFF**	1.180-1.226	0.229-0.251	4-5	0.769-0.897
		SEAL	1.123	0.310	4	0.811
	SHAPE POLAR	**FLUFF**	0.891-1.016	0.400-0.565	3-4	0.787-0.875
		SEAL	1.036	0.450	5	0.925
	ESP POLAR	**FLUFF**	0.891-0.931	0.496-0.538	3-3	0.828-0.829
		SEAL	0.954	0.470	3	0.834
	SHAPE ESP POLAR	**FLUFF**	0.888-0.926	0.501-0.542	3-3	0.833-0.835
		SEAL	0.949	0.476	3	0.840

Table 18. *SADLER* internal validation results. For FLUFF the range (min – max) of values generated by different superpositions is given.

			S_{press}	Q^2	NPC	R^2
SOR	**SHAPE**	**FLUFF**	1.102-1.108	0.212-0.256	-	0.574-0.623
		SEAL	1.098	0.221	-	0.568
	ESP	**FLUFF**	1.167-1.265	0.104-0.153	-	0.380-0.489
		SEAL	1.166	0.153	-	0.488
	POLAR	**FLUFF**	1.103-1.106	0.238-0.243	-	0.399-0.403
		SEAL	1.101	0.246	-	0.405
	SHAPE ESP	**FLUFF**	1.154-1.245	0.135-0.171	-	0.427-0.513
		SEAL	1.153	0.173	-	0.512
	SHAPE POLAR	**FLUFF**	1.099-1.102	0.245-0.249	-	0.413-0.417
		SEAL	1.096	0.252	-	0.418
	ESP POLAR	**FLUFF**	1.062-1.092	0.258-0.298	-	0.464-0.492
		SEAL	1.059	0.303	-	0.495
	SHAPE ESP POLAR	**FLUFF**	1.061-1.089	0.262-0.300	-	0.475-0.500
		SEAL	1.057	0.305	-	0.503
MCSOR	**SHAPE**	**FLUFF**	1.022-1.070	0.409-0.447	6-7	0.991-0.998
		SEAL	1.016	0.398	7	0.998
	ESP	**FLUFF**	0.924-1.271	0.471-0.545	3-5	0.759-0.920
		SEAL	0.953	0.515	5	0.920
	POLAR	**FLUFF**	1.012-1.066	0.445-0.477	6-7	0.941-0.959
		SEAL	1.006	0.483	6	0.940
	SHAPE ESP	**FLUFF**	0.940-1.326	0.441-0.508	4-6	0.925-0.960
		SEAL	0.947	0.501	4	0.924
	SHAPE POLAR	**FLUFF**	0.972-1.010	0.478-0.517	6-6	0.952-0.960
		SEAL	0.984	0.505	6	0.960
	ESP POLAR	**FLUFF**	0.753-1.085	0.412-0.698	5-7	0.942-0.975
		SEAL	0.759	0.693	5	0.954
	SHAPE ESP POLAR	**FLUFF**	0.753-1.072	0.438-0.698	5-7	0.962-0.980
		SEAL	0.762	0.690	5	0.961
SIMPLS	**SHAPE**	**FLUFF**	1.022-1.074	0.436-0.465	6-7	0.991-0.998
		SEAL	1.084	0.425	7	0.998
	ESP	**FLUFF**	0.907-1.253	0.395-0.561	3-5	0.759-0.920
		SEAL	0.934	0.534	5	0.920
	POLAR	**FLUFF**	0.995-1.045	0.465-0.494	6-7	0.941-0.959
		SEAL	0.989	0.499	6	0.940
	SHAPE ESP	**FLUFF**	0.930-1.303	0.432-0.519	4-6	0.925-0.960
		SEAL	0.936	0.513	4	0.924
	SHAPE POLAR	**FLUFF**	0.960-0.997	0.492-0.529	6-6	0.952-0.960
		SEAL	0.972	0.517	6	0.960
	ESP POLAR	**FLUFF**	0.760-1.065	0.432-0.691	5-7	0.942-0.975
		SEAL	0.764	0.688	5	0.954
	SHAPE ESP POLAR	**FLUFF**	0.760-1.052	0.459-0.692	5-7	0.962-0.980
		SEAL	0.767	0.686	5	0.961

Table 19. *EDKB CALF* internal validation results. For FLUFF the range (min – max) of values generated by different superpositions is given.

			S_{press}	Q^2	NPC	R^2
SOR	**SHAPE**	**FLUFF**	0.870-0.875	0.008-0.020	-	0.214-0.226
		SEAL	0.871	0.019	-	0.221
	ESP	**FLUFF**	0.873-0.881	-0.006-0.015	-	0.192-0.258
		SEAL	0.876	0.006	-	0.236
	POLAR	**FLUFF**	0.870-0.895	-0.037-0.021	-	0.209-0.233
		SEAL	0.890	-0.025	-	0.230
	SHAPE ESP	**FLUFF**	0.813-0.865	0.031-0.146	-	0.239-0.384
		SEAL	0.866	0.030	-	0.273
	SHAPE POLAR	**FLUFF**	0.874-0.884	-0.011-0.010	-	0.166-0.184
		SEAL	0.885	-0.013	-	0.167
	ESP POLAR	**FLUFF**	0.830-0.862	0.039-0.108	-	0.297-0.373
		SEAL	0.863	0.037	-	0.332
	SHAPE ESP POLAR	**FLUFF**	0.829-0.860	0.042-0.110	-	0.308-0.386
		SEAL	0.861	0.040	-	0.345
MCSOR	**SHAPE**	**FLUFF**	0.774-0.817	0.221-0.290	3-8	0.682-0.976
		SEAL	0.800	0.221	4	0.800
	ESP	**FLUFF**	0.863-0.951	0.052-0.226	4-15	0.786-0.995
		SEAL	0.905	0.149	11	0.990
	POLAR	**FLUFF**	0.848-0.960	0.101-0.169	6-14	0.846-0.993
		SEAL	0.878	0.100	6	0.871
	SHAPE ESP	**FLUFF**	0.786-0.830	0.179-0.248	4-6	0.730-0.852
		SEAL	0.815	0.191	4	0.749
	SHAPE POLAR	**FLUFF**	0.880-0.965	0.085-0.129	7-15	0.830-0.990
		SEAL	0.896	0.103	8	0.877
	ESP POLAR	**FLUFF**	0.815-0.869	0.119-0.170	2-6	0.414-0.800
		SEAL	0.823	0.140	2	0.442
	SHAPE ESP POLAR	**FLUFF**	0.814-0.854	0.149-0.199	2-6	0.456-0.816
		SEAL	0.822	0.142	2	0.459
SIMPLS	**SHAPE**	**FLUFF**	0.817-0.890	0.184-0.236	4-14	0.834-0.992
		SEAL	0.818	0.185	4	0.834
	ESP	**FLUFF**	0.800-0.841	0.157-0.221	4-6	0.681-0.818
		SEAL	0.821	0.180	4	0.734
	POLAR	**FLUFF**	0.859-0.938	0.055-0.117	2-11	0.272-0.948
		SEAL	0.937	0.086	11	0.939
	SHAPE ESP	**FLUFF**	0.784-0.826	0.185-0.251	4-6	0.730-0.852
		SEAL	0.812	0.197	4	0.749
	SHAPE POLAR	**FLUFF**	0.878-0.963	0.089-0.134	7-15	0.830-0.990
		SEAL	0.894	0.107	8	0.877
	ESP POLAR	**FLUFF**	0.813-0.864	0.128-0.177	2-6	0.414-0.800
		SEAL	0.821	0.145	2	0.442
	SHAPE ESP POLAR	**FLUFF**	0.812-0.850	0.156-0.205	2-6	0.456-0.816
		SEAL	0.820	0.146	2	0.459

Table 20. *EDKB HUMANA* internal validation results. For FLUFF the range (min – max) of values generated by different superpositions is given.

			S_{press}	Q^2	NPC	R^2
SOR	**SHAPE**	**FLUFF**	1.318-1.351	0.092-0.137	-	0.447-0.503
		SEAL	1.351	0.091	-	0.447
	ESP	**FLUFF**	1.216-1.249	0.224-0.265	-	0.646-0.675
		SEAL	1.251	0.222	-	0.667
	POLAR	**FLUFF**	1.150-1.166	0.324-0.342	-	0.477-0.485
		SEAL	1.156	0.335	-	0.480
	SHAPE ESP	**FLUFF**	1.390-1.427	-0.013-0.039	-	0.321-0.449
		SEAL	1.408	0.014	-	0.375
	SHAPE POLAR	**FLUFF**	1.218-1.233	0.244-0.262	-	0.334-0.351
		SEAL	1.224	0.255	-	0.345
	ESP POLAR	**FLUFF**	1.209-1.235	0.241-0.272	-	0.375-0.398
		SEAL	1.216	0.264	-	0.392
	SHAPE ESP POLAR	**FLUFF**	1.204-1.229	0.249-0.279	-	0.393-0.413
		SEAL	1.211	0.270	-	0.408
MCSOR	**SHAPE**	**FLUFF**	1.042-1.181	0.442-0.484	3-13	0.908-0.998
		SEAL	1.045	0.475	3	0.919
	ESP	**FLUFF**	1.063-1.196	0.445-0.515	4-15	0.920-0.998
		SEAL	1.096	0.444	5	0.968
	POLAR	**FLUFF**	0.994-1.061	0.538-0.597	7-15	0.978-1.000
		SEAL	1.042	0.588	15	0.999
	SHAPE ESP	**FLUFF**	1.219-1.266	0.270-0.330	6-7	0.946-0.966
		SEAL	1.239	0.301	6	0.946
	SHAPE POLAR	**FLUFF**	0.994-1.025	0.528-0.540	4-7	0.850-0.955
		SEAL	1.025	0.530	7	0.954
	ESP POLAR	**FLUFF**	1.122-1.191	0.378-0.439	6-9	0.919-0.972
		SEAL	1.251	0.340	10	0.971
	SHAPE ESP POLAR	**FLUFF**	1.083-1.160	0.410-0.467	6-9	0.935-0.979
		SEAL	1.169	0.389	7	0.955
SIMPLS	**SHAPE**	**FLUFF**	1.123-1.156	0.375-0.410	2-7	0.897-0.982
		SEAL	1.156	0.403	7	0.981
	ESP	**FLUFF**	1.265-1.331	0.208-0.267	5-7	0.870-0.951
		SEAL	1.291	0.227	5	0.872
	POLAR	**FLUFF**	1.013-1.113	0.493-0.541	6-13	0.860-0.991
		SEAL	1.028	0.528	7	0.930
	SHAPE ESP	**FLUFF**	1.218-1.265	0.271-0.330	6-7	0.946-0.966
		SEAL	1.238	0.302	6	0.946
	SHAPE POLAR	**FLUFF**	0.993-1.024	0.529-0.541	4-7	0.850-0.955
		SEAL	1.024	0.531	7	0.954
	ESP POLAR	**FLUFF**	1.120-1.187	0.383-0.442	6-9	0.919-0.972
		SEAL	1.246	0.346	10	0.971
	SHAPE ESP POLAR	**FLUFF**	1.082-1.156	0.414-0.469	5-9	0.901-0.979
		SEAL	1.165	0.393	7	0.955

Table 21. *EDKB HUMANB* internal validation results. For FLUFF the range (min – max) of values generated by different superpositions is given.

			S_{press}	Q^2	NPC	R^2
SOR	**SHAPE**	**FLUFF**	1.195-1.208	0.160-0.178	-	0.397-0.432
		SEAL	1.208	0.160	-	0.397
	ESP	**FLUFF**	1.175-1.194	0.179-0.205	-	0.561-0.596
		SEAL	1.190	0.185	-	0.584
	POLAR	**FLUFF**	1.088-1.100	0.303-0.319	-	0.455-0.463
		SEAL	1.094	0.311	-	0.458
	SHAPE ESP	**FLUFF**	1.230-1.312	0.008-0.129	-	0.266-0.367
		SEAL	1.274	0.065	-	0.315
	SHAPE POLAR	**FLUFF**	1.164-1.176	0.204-0.219	-	0.321-0.336
		SEAL	1.171	0.210	-	0.328
	ESP POLAR	**FLUFF**	1.133-1.174	0.207-0.260	-	0.386-0.432
		SEAL	1.145	0.245	-	0.410
	SHAPE ESP POLAR	**FLUFF**	1.127-1.166	0.217-0.269	-	0.407-0.451
		SEAL	1.138	0.254	-	0.430
MCSOR	**SHAPE**	**FLUFF**	1.081-1.201	0.316-0.350	3-12	0.861-0.996
		SEAL	1.090	0.339	3	0.861
	ESP	**FLUFF**	1.052-1.084	0.366-0.417	3-6	0.862-0.966
		SEAL	1.068	0.366	3	0.862
	POLAR	**FLUFF**	1.012-1.114	0.409-0.472	6-15	0.948-0.999
		SEAL	1.049	0.420	6	0.954
	SHAPE ESP	**FLUFF**	1.177-1.245	0.168-0.255	3-5	0.766-0.893
		SEAL	1.198	0.202	3	0.766
	SHAPE POLAR	**FLUFF**	0.995-1.066	0.405-0.459	4-7	0.817-0.927
		SEAL	1.052	0.406	5	0.849
	ESP POLAR	**FLUFF**	1.074-1.195	0.247-0.381	2-6	0.538-0.896
		SEAL	1.145	0.245	1	0.410
	SHAPE ESP POLAR	**FLUFF**	1.036-1.180	0.267-0.424	4-6	0.840-0.916
		SEAL	1.183	0.262	6	0.891
SIMPLS	**SHAPE**	**FLUFF**	1.112-1.134	0.272-0.301	2-2	0.824-0.843
		SEAL	1.113	0.298	2	0.824
	ESP	**FLUFF**	1.241-1.320	0.065-0.152	3-6	0.659-0.896
		SEAL	1.257	0.121	3	0.661
	POLAR	**FLUFF**	0.995-1.054	0.426-0.478	6-7	0.832-0.894
		SEAL	1.057	0.422	7	0.892
	SHAPE ESP	**FLUFF**	1.176-1.242	0.172-0.258	3-5	0.766-0.893
		SEAL	1.196	0.205	3	0.766
	SHAPE POLAR	**FLUFF**	0.994-1.064	0.407-0.460	4-7	0.817-0.927
		SEAL	1.051	0.407	5	0.849
	ESP POLAR	**FLUFF**	1.070-1.197	0.251-0.385	2-9	0.538-0.956
		SEAL	1.145	0.246	1	0.410
	SHAPE ESP POLAR	**FLUFF**	1.034-1.176	0.271-0.426	4-6	0.840-0.916
		SEAL	1.179	0.267	6	0.891

Table 22. *EDKB MOUSE* internal validation results. For FLUFF the range (min – max) of values generated by different superpositions is given.

			S_{press}	Q^2	NPC	R^2
SOR	**SHAPE**	**FLUFF**	1.377-1.390	0.286-0.300	-	0.491-0.503
		SEAL	1.388	0.288	-	0.491
	ESP	**FLUFF**	1.350-1.373	0.304-0.327	-	0.503-0.525
		SEAL	1.373	0.304	-	0.504
	POLAR	**FLUFF**	1.272-1.283	0.392-0.403	-	0.513-0.525
		SEAL	1.283	0.392	-	0.513
	SHAPE ESP	**FLUFF**	1.465-1.501	0.168-0.208	-	0.324-0.344
		SEAL	1.501	0.168	-	0.324
	SHAPE POLAR	**FLUFF**	1.323-1.331	0.345-0.353	-	0.428-0.436
		SEAL	1.331	0.346	-	0.428
	ESP POLAR	**FLUFF**	1.233-1.264	0.410-0.438	-	0.532-0.554
		SEAL	1.264	0.410	-	0.532
	SHAPE ESP POLAR	**FLUFF**	1.230-1.261	0.413-0.441	-	0.542-0.564
		SEAL	1.261	0.413	-	0.542
MCSOR	**SHAPE**	**FLUFF**	1.281-1.303	0.420-0.439	6-6	0.953-0.963
		SEAL	1.303	0.420	6	0.953
	ESP	**FLUFF**	1.307-1.392	0.304-0.413	1-7	0.503-0.966
		SEAL	1.373	0.304	1	0.504
	POLAR	**FLUFF**	1.202-1.294	0.431-0.514	5-7	0.929-0.965
		SEAL	1.280	0.431	5	0.929
	SHAPE ESP	**FLUFF**	1.347-1.414	0.339-0.380	6-8	0.887-0.935
		SEAL	1.424	0.329	8	0.929
	SHAPE POLAR	**FLUFF**	1.307-1.341	0.376-0.406	5-5	0.848-0.869
		SEAL	1.341	0.376	5	0.848
	ESP POLAR	**FLUFF**	1.164-1.353	0.410-0.552	1-10	0.532-0.959
		SEAL	1.264	0.410	1	0.532
	SHAPE ESP POLAR	**FLUFF**	1.167-1.336	0.419-0.549	8-9	0.936-0.965
		SEAL	1.340	0.416	9	0.952
SIMPLS	**SHAPE**	**FLUFF**	1.267-1.384	0.418-0.442	5-13	0.943-0.988
		SEAL	1.295	0.418	5	0.943
	ESP	**FLUFF**	1.391-1.449	0.282-0.351	6-8	0.835-0.911
		SEAL	1.449	0.282	6	0.841
	POLAR	**FLUFF**	1.330-1.353	0.342-0.371	1-5	0.414-0.800
		SEAL	1.335	0.342	1	0.414
	SHAPE ESP	**FLUFF**	1.346-1.413	0.340-0.381	6-8	0.887-0.935
		SEAL	1.422	0.331	8	0.929
	SHAPE POLAR	**FLUFF**	1.304-1.337	0.379-0.410	5-5	0.848-0.869
		SEAL	1.337	0.379	5	0.848
	ESP POLAR	**FLUFF**	1.162-1.348	0.410-0.554	1-10	0.532-0.959
		SEAL	1.264	0.410	1	0.532
	SHAPE ESP POLAR	**FLUFF**	1.165-1.331	0.424-0.551	8-9	0.936-0.965
		SEAL	1.335	0.420	9	0.952

Table 23. *EDKB RAT* internal validation results. For FLUFF the range (min – max) of values generated by different superpositions is given.

			S_{press}	Q^2	NPC	R^2
SOR	**SHAPE**	**FLUFF**	1.538-1.570	0.229-0.260	-	0.459-0.478
		SEAL	1.539	0.259	-	0.476
	ESP	**FLUFF**	1.478-1.532	0.266-0.316	-	0.447-0.495
		SEAL	1.484	0.311	-	0.492
	POLAR	**FLUFF**	1.462-1.473	0.321-0.332	-	0.406-0.421
		SEAL	1.469	0.325	-	0.416
	SHAPE ESP	**FLUFF**	1.728-1.735	0.059-0.066	-	0.185-0.212
		SEAL	1.735	0.059	-	0.202
	SHAPE POLAR	**FLUFF**	1.523-1.536	0.262-0.274	-	0.335-0.342
		SEAL	1.536	0.262	-	0.337
	ESP POLAR	**FLUFF**	1.491-1.499	0.297-0.304	-	0.392-0.398
		SEAL	1.502	0.295	-	0.390
	SHAPE ESP POLAR	**FLUFF**	1.485-1.491	0.305-0.310	-	0.409-0.414
		SEAL	1.494	0.302	-	0.407
MCSOR	**SHAPE**	**FLUFF**	1.305-1.392	0.437-0.480	4-11	0.876-0.991
		SEAL	1.313	0.474	4	0.877
	ESP	**FLUFF**	1.262-1.351	0.442-0.517	4-5	0.868-0.906
		SEAL	1.266	0.514	5	0.905
	POLAR	**FLUFF**	1.255-1.373	0.438-0.519	4-7	0.870-0.967
		SEAL	1.264	0.512	4	0.875
	SHAPE ESP	**FLUFF**	1.292-1.389	0.448-0.506	8-12	0.958-0.986
		SEAL	1.375	0.446	9	0.957
	SHAPE POLAR	**FLUFF**	1.336-1.399	0.407-0.503	5-15	0.752-0.994
		SEAL	1.356	0.488	15	0.994
	ESP POLAR	**FLUFF**	1.262-1.401	0.438-0.541	9-15	0.944-0.987
		SEAL	1.353	0.472	11	0.964
	SHAPE ESP POLAR	**FLUFF**	1.272-1.419	0.429-0.529	10-13	0.962-0.985
		SEAL	1.339	0.479	10	0.971
SIMPLS	**SHAPE**	**FLUFF**	1.308-1.421	0.394-0.482	3-7	0.848-0.976
		SEAL	1.308	0.473	3	0.864
	ESP	**FLUFF**	1.380-1.507	0.334-0.441	6-12	0.843-0.967
		SEAL	1.506	0.335	9	0.914
	POLAR	**FLUFF**	1.303-1.352	0.446-0.519	5-13	0.633-0.983
		SEAL	1.318	0.507	13	0.983
	SHAPE ESP	**FLUFF**	1.291-1.387	0.449-0.507	8-12	0.958-0.986
		SEAL	1.374	0.446	9	0.957
	SHAPE POLAR	**FLUFF**	1.334-1.397	0.409-0.504	5-15	0.752-0.994
		SEAL	1.354	0.489	15	0.994
	ESP POLAR	**FLUFF**	1.260-1.398	0.441-0.542	9-15	0.944-0.987
		SEAL	1.351	0.474	11	0.964
	SHAPE ESP POLAR	**FLUFF**	1.270-1.416	0.432-0.531	10-13	0.962-0.985
		SEAL	1.336	0.481	10	0.971

4.4 MCSOR vis-à-vis other PLS methods

When introducing a new multivariate regression technique, one fundamental question immediately arises: How will it perform in comparison to the more established multivariate methods? In order to evaluate this, the MCSOR should be compared directly with other techniques. Unfortunately there is a veritable cornucopia of different methodologies to choose from and an exhaustive evaluation of the relative performances is virtually impossible. However, in recent years, PLS has become de facto basic tool of chemometrics[346], and thus it provides a suitable frame of reference for the performance of MCSOR.

In this validation work the MCSOR was used to build a model of five different data sets (13 counting the subsets) for which the performance of PLS was known or expected to be good. In order to reliably evaluate the relative performance of MCSOR, the same data were also analysed with three variants of the widely employed partial least-squares (PLS) regression methods, namely SIMPLS[350], SVDPLS[352] and PPLS[508], which act as reference techniques. In general, it should be emphasised that the primary aim of the validation was not to develop alternatives for the original models, but to compare the performance and validity of different prediction algorithms.

4.4.1 Experimental data and variable selection

The first two datasets have already been introduced, as they are the CBG and TBG sets originally used by Cramer[14]. In this work the EVA[270] spectroscopic descriptors, computed as described in previous works[272,509], were used. The third data set comes with the DRAGON software[96] and it consists of 42 organic molecules for which the melting point (MP) and boiling point (BP) are known. When the DRAGON descriptors were computed and autoscaled, after which zero or constant descriptors, as well as descriptors without a strong correlation (|correlation coefficient| < 0.75), were excluded. This left a set of 130 descriptors for the MP data set and 203 for BP data set, respectively.

The fourth set (SUGAR) is a multivariate calibration set which contains 125 records of near-infrared (NIR) absorbance spectra of mixtures of three sugars: sucrose, glucose and fructose in aqueous solution, each at 5 levels (6, 10, 12, 14, and 18 percent by mass) in a full $5^3 = 125$ experimental design. This set is described in detail by Brown et al [510], and is also freely available on the Internet (http://www.blackwellpublishing.com/rss/Volumes/Bv64p3_read1.htm). The fifth and last set (DIESEL) contains the NIR spectra of diesel fuels along with various properties of those fuels including boiling points (at 50% recovery, deg C), cetane number (similar to octane number but for diesel), density (g/mL at15 deg C), freezing temperature of the fuel (deg C), total aromatics (mass%) and viscosity (cSt at 40 deg C). This data set was originally provided by S. Hutzler, of Southwest Research Institute at San Antonio, TX, USA, and is also available on the Internet at http://software.eigenvector.com/Data/SWRI/index.html.

Variable selection (reduction) is a critical issue for all multivariate methods. In multivariate data most of the X-vectors contain at least some information about Y, but there are usually some 20-30% of the variables which have less information than noise. If such variables can be reliably

identified, they should be deleted. Also, it is often possible to reduce the number of variables further without any apparent decrease in fit. This is due to the fact that the variables are often strongly correlated and the removal of variables does not lead to a significant loss of information. In doing so, however, the role of the remaining variables will be overemphasised, and a bias is introduced. Thus it seems logical to keep all variables with reasonable level of correlation in order to guarantee the maximal stability and predictive ability of the models. Therefore, no variable selection was applied in this study, except for the MP/BP data sets, where the improvement was substantial. It is possible that the variables with low signal to noise ratio are somewhat detrimental for both PLS and MCSOR. The MCSOR algorithm however, tends to discard unimportant (i.e. non-predictive) variables which have a negligible contribution to the final models, and thus the variable selection is perhaps not so crucial for MCSOR as is the case with many other multivariate methods.

4.4.2 Statistical methods and model validation

For each data set, a large number of PLS and MCSOR models, 500 in all, were derived by choosing 2/3 samples randomly for the training set and placing the remaining 1/3 in the test set. The maximum number of components allowed was selected using the one quarter rule. Smaller CBG and TBG sets were limited to 5 and 4 components, respectively. Larger BP and MP sets got at most 7 components, whereas for SUGAR and DIESEL data the maximum number of components was 20 and 25, respectively.

The internal predictability of each model was assessed by leave-one-out cross-validation (LOO CV), and the optimum number of (principal, in PLS) components was selected on the basis of the maximum Q2. Standard statistical indicators of internal predictivity, namely SPRESS and Q2, were evaluated for all of the 500 models. To assess the external predictability, the conventional correlation coefficient (R2ex), mean absolute deviation (|Δ|ave), standard error of prediction (SDEP) and predictive r2-score (Pr-r2) were computed (For more details on statistical descriptors, see Section 2.5).

4.4.3 Comparison of MCSOR and PLS performances

When the results of the five datasets are compared it seems that the S_{press} and Q^2 values in internal (LOO CV) predictions are usually slightly better with MCSOR than those with PLS which may be due to fact that in MCSOR the Y-vector is deflated and X is kept unchanged. In external (LMO CV) predictions, the $|\Delta|_{ave}$, SDEP and Pr-r^2 values are also better or at least equally good to those of different PLS algorithms. In general, the differences in performance between different PLS algorithms can be surprisingly large, whereas MCSOR is stable throughout the tests.

For CBG data (Table 24), all methods worked quite well, both in internal and external predictions. Still, MCSOR is slightly better than PLS, and there are some differences in the performances of PLS algorithms. For TBG data set (Table 24) this trend is amplified as SVDPLS and PPLS disqualify badly. It should be emphasised that CBG and TBG models derived using MCSOR are highly predictive also with the scaled X-block data, whereas the scaling of the

EVA descriptors is usually detrimental for PLS [101,102,269,272]. For the prediction of DRAGON MPs, MCSOR performs best, SIMPLS is slightly inferior, and once again SVDPLS and PPLS disqualify clearly (Table 25). For the BP data, MCSOR performs best again, but the SVDPLS is now superior to SIMPLS and PPLS, which yield nearly equal performance (Table 25). For SUGAR data (Table 26), MCSOR and SVDPLS perform best, SIMPLS is slightly inferior, and PPLS disqualifies clearly. For DIESEL data (Table 27 and Table 28), the performances of MCSOR, SIMPLS and SVDPLS are nearly equal, and PPLS is only slightly inferior to them. It should be emphasised that even though all methods take a large number of components to explain SUGAR and DIESEL data this does not seem to disturb external predictions.

Table 24. Statistical performance indicators for MCSOR and PLS with CBG and TBG EVA data. Values are averages (± std, min - max) over 500 randomized runs. The sizes of training and test sets were 21/10 and 14/7 for CBG and TBG, respectively.

	SPRESS	Q^2	R^2_{ex}	$\|\Delta\|_{ave}$	SDEP	Pr-R^2	NPC
CBG MCSOR	0.64 ±0.06 (0.46 - 0.83)	0.71 ±0.07 (0.45 - 0.88)	0.76 ±0.11 (0.21 - 0.96)	0.48 ±0.10 (0.26 - 0.94)	0.57 ±0.11 (0.29 - 1.06)	0.72 ±0.10 (0.19 - 0.91)	3.61 ±1.04 (2 - 5)
CBG SIMPLS	0.66 ±0.06 (0.49 - 0.89)	0.69 ±0.07 (0.40 - 0.86)	0.77 ±0.11 (0.21 - 0.96)	0.49 ±0.11 (0.24 - 0.93)	0.58 ±0.12 (0.27 - 1.07)	0.71 ±0.10 (0.32 - 0.92)	3.64 ±1.01 (2 - 5)
CBG SVDPLS	0.71 ±0.07 (0.50 - 0.90)	0.66 ±0.08 (0.40 - 0.82)	0.71 ±0.15 (0.23 - 0.94)	0.50 ±0.11 (0.23 - 0.87)	0.61 ±0.14 (0.27 - 1.04)	0.68 ±0.14 (0.13 - 0.94)	4.54 ±0.68 (2 - 5)
CBG PPLS	0.79 ±0.09 (0.53 - 1.02)	0.57 ±0.10 (0.18 - 0.82)	0.63 ±0.18 (0.11 - 0.96)	0.57 ±0.14 (0.24 - 1.10)	0.69 ±0.15 (0.31 - 1.19)	0.59 ±0.18 (-0.19 - 0.92)	4.38 ±0.90 (1 - 5)
TBG MCSOR	1.06 ±0.15 (0.59 - 1.46)	0.43 ±0.18 (-0.28 - 0.80)	0.57 ±0.25 (0.00 - 0.97)	0.69 ±0.22 (0.18 - 1.66)	0.85 ±0.24 (0.23 - 1.89)	0.48 ±0.28 (-1.84 - 0.96)	3.64 ±0.71 (1 - 4)
TBG SIMPLS	1.08 ±0.14 (0.64 - 1.42)	0.41 ±0.17 (-0.26 - 0.76)	0.57 ±0.26 (0.00 - 0.97)	1.03 ±0.45 (0.23 - 2.85)	1.29 ±0.37 (0.28 - 4.71)	0.36 ±0.31 (-3.72 - 0.95)	3.64 ±0.71 (1 - 4)
TBG SVDPLS	1.35 ±0.19 (0.81 - 1.88)	0.07 ±0.31 (-1.38 - 0.69)	0.40 ±0.27 (0.00 - 0.95)	1.52 ±0.56 (0.26 - 8.81)	1.44 ±0.58 (0.31 - 8.46)	0.08 ±0.48 (-4.81 - 0.88)	3.83 ±0.54 (1 - 4)
TBG PPLS	1.58 ±0.23 (0.87 - 2.25)	-0.29 ±0.39 (-2.32 - 0.60)	0.26 ±0.23 (0.00 - 0.89)	1.11 ±0.35 (0.40 - 2.21)	1.33 ±0.39 (0.48 - 2.58)	0.13 ±0.39 (-3.51 - 0.72)	3.56 ±0.71 (1 - 4)

Table 25. Statistical performance indicators for MCSOR and PLS with DRAGON MP and BP data. Values are averages (± std, min - max) over 500 randomized runs. The sizes of training and test sets were 28/14, respectively.

	SPRESS	Q^2	R^2_{ex}	$\|\Delta\|_{ave}$	SDEP	Pr-R^2	NPC
MP MCSOR	23.31 ±4.99 (11.62 - 46.01)	0.96 ±0.02 (0.84 - 0.98)	0.97 ±0.03 (0.79 - 1.00)	13.18 ±5.36 (4.10 - 39.64)	17.70 ±7.25 (4.58 - 58.12)	0.96 ±0.04 (0.64 - 1.00)	7.00 ±0.06 (6 - 7)
MP SIMPLS	26.58 ±5.44 (15.29 - 47.25)	0.94 ±0.03 (0.71 - 0.98)	0.97 ±0.03 (0.79 - 1.00)	16.14 ±6.23 (4.97 - 49.82)	20.53 ±7.61 (6.06 - 62.34)	0.95 ±0.04 (0.60 - 1.00)	6.99 ±0.08 (6 - 7)
MP SVDPLS	36.20 ±6.56 (13.74 - 43.49)	0.92 ±0.15 (0.35 - 0.90)	0.69 ±0.16 (0.01 - 0.95)	33.67 ±18.60 (5.33 - 76.85)	42.98 ±8.39 (6.68 - 72.93)	0.93 ±0.21 (-1.66 - 0.55)	6.59 ±0.36 (4 - 7)
MP PPLS	36.49 ±6.47 (24.39 - 45.57)	0.51 ±0.15 (0.31 - 0.78)	0.76 ±0.13 (0.28 - 0.98)	59.46 ±9.69 (3.48 - 116.86)	55.03 ±13.21 (41.47 - 152.96)	0.80 ±0.33 (-1.56 - 0.85)	5.42 ±0.54 (4 - 7)
BP MCSOR	42.47 ±5.80 (25.57 - 62.56)	0.88 ±0.04 (0.74 - 0.95)	0.88 ±0.08 (0.41 - 0.99)	31.14 ±7.64 (15.05 - 59.44)	38.23 ±10.19 (19.99 - 81.98)	0.86 ±0.09 (0.13 - 0.97)	6.27 ±1.33 (1 - 7)
BP SIMPLS	44.77 ±6.15 (30.59 - 63.09)	0.87 ±0.04 (0.70 - 0.95)	0.88 ±0.08 (0.41 - 0.99)	45.31 ±11.86 (16.93 - 116.56)	55.41 ±14.88 (21.89 - 143.33)	0.81 ±0.10 (-1.05 - 0.97)	6.22 ±1.37 (1 - 7)
BP SVDPLS	57.61 ±7.08 (47.97 - 67.46)	0.77 ±0.21 (-0.92 - 0.94)	0.79 ±0.12 (0.00 - 0.97)	37.69 ±17.68 (15.64 - 124.80)	47.49 ±23.04 (17.78 - 158.88)	0.85 ±0.32 (-2.02 - 0.94)	6.08 ±0.63 (1 - 7)
BP PPLS	57.37 ±7.50 (47.45 - 69.37)	0.72 ±0.20 (-0.92 - 0.96)	0.80 ±0.11 (0.30 - 0.97)	50.72 ±11.48 (7.47 - 141.74)	56.23 ±13.23 (40.28 - 159.88)	0.73 ±0.48 (-3.28 - 0.85)	5.91 ±1.21 (1 - 7)

Table 26. Statistical performance indicators for MCSOR and PLS with SUGAR NIR data. Values are averages (± std, min - max) over 500 randomized runs. The sizes of training and test sets were 80/45, respectively.

	SPRESS	Q^2	R^2_{ex}	$\|\Delta\|_{ave}$	SDEP	Pr-R^2	NPC
SUCROSE MCSOR	0.36 ±0.04 (0.24 - 0.49)	0.99 ±0.00 (0.99 - 1.00)	0.99 ±0.00 (0.96 - 1.00)	0.24 ±0.03 (0.16 - 0.45)	0.45 ±0.09 (0.27 - 1.07)	0.99 ±0.00 (0.95 - 1.00)	18.51 ±1.23 (14 - 20)
SUCROSE SIMPLS_MC	0.41 ±0.12 (0.22 - 1.02)	0.99 ±0.01 (0.95 - 1.00)	0.99 ±0.00 (0.98 - 1.00)	0.32 ±0.12 (0.17 - 1.00)	0.54 ±0.18 (0.27 - 1.43)	0.99 ±0.01 (0.94 - 1.00)	12.27 ±4.16 (6 - 20)
SUCROSE SVDPLS	0.36 ±0.04 (0.25 - 0.48)	0.99 ±0.00 (0.99 - 1.00)	0.99 ±0.00 (0.96 - 1.00)	0.24 ±0.04 (0.17 - 0.42)	0.45 ±0.09 (0.27 - 0.98)	0.99 ±0.00 (0.96 - 1.00)	19.14 ±1.13 (15 - 20)
SUCROSE PPLS	1.66 ±0.42 (0.61 - 2.80)	0.86 ±0.07 (0.63 - 0.98)	0.87 ±0.13 (0.44 - 1.00)	0.98 ±0.59 (0.15 - 2.30)	1.79 ±1.07 (0.27 - 4.56)	0.85 ±0.15 (0.30 - 1.00)	19.71 ±1.42 (10 - 20)
GLUCOSE MCSOR	0.44 ±0.07 (0.22 - 0.66)	0.99 ±0.00 (0.98 - 1.00)	0.99 ±0.00 (0.95 - 1.00)	0.27 ±0.05 (0.15 - 0.48)	0.54 ±0.09 (0.27 - 1.16)	0.99 ±0.01 (0.95 - 1.00)	14.91 ±2.16 (10 - 20)
GLUCOSE SIMPLS	0.50 ±0.10 (0.23 - 0.95)	0.99 ±0.01 (0.94 - 1.00)	0.99 ±0.00 (0.98 - 1.00)	0.37 ±0.11 (0.20 - 0.81)	0.63 ±0.18 (0.36 - 1.16)	0.98 ±0.01 (0.93 - 1.00)	11.73 ±4.33 (6 - 20)
GLUCOSE SVDPLS	0.47 ±0.08 (0.24 - 0.66)	0.99 ±0.00 (0.97 - 1.00)	0.99 ±0.00 (0.95 - 1.00)	0.29 ±0.05 (0.16 - 0.53)	0.63 ±0.09 (0.27 - 1.34)	0.99 ±0.00 (0.95 - 1.00)	17.05 ±2.26 (11 - 20)
GLUCOSE PPLS	0.80 ±0.11 (0.46 - 1.25)	0.97 ±0.01 (0.93 - 0.99)	0.97 ±0.01 (0.82 - 0.99)	0.52 ±0.11 (0.28 - 1.26)	0.89 ±0.18 (0.54 - 2.06)	0.97 ±0.02 (0.79 - 0.99)	19.46 ±0.91 (15 - 20)
FRUCTOSE MCSOR	0.47 ±0.06 (0.30 - 0.81)	0.99 ±0.00 (0.97 - 1.00)	0.99 ±0.00 (0.95 - 1.00)	0.31 ±0.05 (0.19 - 0.61)	0.63 ±0.09 (0.36 - 1.52)	0.99 ±0.01 (0.95 - 1.00)	16.69 ±1.95 (12 - 20)
FRUCTOSE SIMPLS	0.47 ±0.11 (0.27 - 1.03)	0.99 ±0.01 (0.94 - 1.00)	0.99 ±0.00 (0.98 - 1.00)	0.35 ±0.12 (0.17 - 0.98)	0.63 ±0.18 (0.27 - 1.34)	0.99 ±0.01 (0.93 - 1.00)	8.43 ±2.32 (5 - 20)
FRUCTOSE SVDPLS	0.46 ±0.06 (0.29 - 0.79)	0.99 ±0.00 (0.97 - 1.00)	0.99 ±0.00 (0.96 - 1.00)	0.30 ±0.05 (0.18 - 0.62)	0.54 ±0.09 (0.36 - 1.61)	0.99 ±0.00 (0.95 - 1.00)	17.35 ±1.67 (13 - 20)
FRUCTOSE PPLS	1.21 ±0.26 (0.55 - 1.98)	0.93 ±0.03 (0.80 - 0.98)	0.93 ±0.06 (0.70 - 0.99)	0.73 ±0.35 (0.23 - 1.75)	1.34 ±0.63 (0.45 - 4.11)	0.93 ±0.07 (0.65 - 0.99)	19.90 ±0.46 (16 - 20)

Table 27. Statistical performance indicators for MCSOR and PLS with DIESEL NIR data. Values are averages (± std, min - max) over 500 randomized runs. The sizes of training and test sets were 100/295, respectively.

	SPRESS	Q^2	R^2_{ex}	$\|\Delta\|_{ave}$	SDEP	Pr-R^2	NPC
BOILING POINT MCSOR	5.32 ±0.92 (3.28 - 9.78)	0.94 ±0.02 (0.83 - 0.98)	0.93 ±0.03 (0.70 - 0.97)	3.78 ±0.44 (2.96 - 6.19)	3.30 ±0.70 (2.40 - 7.80)	0.92 ±0.03 (0.70 - 0.97)	19.14 ±3.79 (7 - 25)
BOILING POINT SIMPLS	5.26 ±0.86 (3.16 - 8.53)	0.95 ±0.02 (0.83 - 0.98)	0.93 ±0.05 (0.67 - 0.97)	3.83 ±0.56 (2.74 - 6.17)	3.40 ±0.90 (2.30 - 8.00)	0.92 ±0.05 (0.66 - 0.97)	18.32 ±4.11 (6 - 25)
BOILING POINT SVDPLS	5.28 ±0.91 (3.20 - 9.47)	0.95 ±0.02 (0.82 - 0.98)	0.93 ±0.03 (0.73 - 0.97)	3.75 ±0.44 (2.99 - 6.26)	3.30 ±0.60 (2.40 - 6.90)	0.93 ±0.03 (0.72 - 0.97)	19.19 ±3.15 (8 - 25)
BOILING POINT PPLS	8.76 ±1.03 (6.03 - 14.11)	0.86 ±0.04 (0.65 - 0.95)	0.84 ±0.04 (0.68 - 0.94)	6.16 ±0.77 (3.80 - 9.07)	4.80 ±0.70 (3.00 - 7.70)	0.83 ±0.05 (0.59 - 0.94)	22.64 ±3.83 (7 - 25)
CETANE NUMBER MCSOR	2.46 ±0.20 (1.84 - 2.96)	0.54 ±0.09 (0.23 - 0.77)	0.52 ±0.05 (0.32 - 0.65)	1.91 ±0.11 (1.69 - 2.40)	1.50 ±0.10 (1.30 - 2.20)	0.30 ±0.19 (-0.57 - 0.57)	8.87 ±3.21 (3 - 18)
CETANE NUMBER SIMPLS	2.47 ±0.21 (1.89 - 3.15)	0.54 ±0.09 (0.20 - 0.75)	0.52 ±0.06 (0.19 - 0.63)	1.90 ±0.12 (1.62 - 2.49)	1.50 ±0.10 (1.30 - 2.70)	0.29 ±0.22 (-2.65 - 0.58)	9.20 ±3.42 (2 - 23)
CETANE NUMBER SVDPLS	2.50 ±0.19 (1.91 - 3.08)	0.53 ±0.08 (0.21 - 0.78)	0.51 ±0.05 (0.26 - 0.62)	1.94 ±0.11 (1.73 - 2.69)	1.50 ±0.10 (1.30 - 2.20)	0.27 ±0.19 (-0.70 - 0.58)	9.41 ±3.36 (4 - 24)
CETANE NUMBER PPLS	2.59 ±0.21 (2.00 - 3.17)	0.51 ±0.08 (0.22 - 0.78)	0.50 ±0.04 (0.32 - 0.59)	1.97 ±0.11 (1.72 - 2.45)	1.50 ±0.10 (1.30 - 2.00)	0.20 ±0.20 (-0.56 - 0.55)	12.12 ±5.60 (4 - 25)
DENSITY MCSOR	0.00 ±0.00 (0.00 - 0.00)	0.99 ±0.00 (0.95 - 1.00)	0.99 ±0.01 (0.88 - 0.99)	0.00 ±0.00 (0.00 - 0.00)	0.00 ±0.00 (0.00 - 0.00)	0.99 ±0.01 (0.88 - 0.99)	22.47 ±2.92 (9 - 25)
DENSITY SIMPLS	0.00 ±0.00 (0.00 - 0.01)	0.98 ±0.01 (0.81 - 1.00)	0.99 ±0.01 (0.84 - 1.00)	0.00 ±0.00 (0.00 - 0.00)	0.00 ±0.00 (0.00 - 0.00)	0.98 ±0.02 (0.84 - 1.00)	21.61 ±3.65 (11 - 25)
DENSITY SVDPLS	0.00 ±0.00 (0.00 - 0.00)	0.99 ±0.01 (0.94 - 1.00)	0.98 ±0.01 (0.91 - 0.99)	0.00 ±0.00 (0.00 - 0.00)	0.00 ±0.00 (0.00 - 0.00)	0.98 ±0.01 (0.90 - 0.99)	21.91 ±2.91 (12 - 25)
DENSITY PPLS	0.00 ±0.00 (0.00 - 0.01)	0.93 ±0.03 (0.79 - 0.98)	0.93 ±0.03 (0.71 - 0.98)	0.00 ±0.00 (0.00 - 0.00)	0.00 ±0.00 (0.00 - 0.00)	0.92 ±0.03 (0.70 - 0.98)	23.88 ±2.26 (11 - 25)

Table 28. Statistical performance indicators for MCSOR and PLS with DIESEL NIR data. Values are averages (± std, min - max) over 500 randomized runs. The sizes of training and test sets were 100/295, respectively.

	SPRESS	Q^2	R^2_{ex}	$\|\Delta\|_{ave}$	SDEP	Pr-R^2	NPC
FREEZING TEMPERATURE MCSOR	5.42 ±0.62 (3.78 - 8.37)	0.80 ±0.07 (0.44 - 0.93)	0.78 ±0.08 (0.13 - 0.85)	4.04 ±0.37 (3.45 - 6.58)	3.30 ±0.70 (2.70 - 11.60)	0.73 ±0.09 (0.04 - 0.84)	11.21 ±4.26 (5 - 25)
FREEZING TEMPERATURE SIMPLS	5.42 ±0.63 (3.87 - 8.05)	0.79 ±0.07 (0.33 - 0.93)	0.78 ±0.08 (0.11 - 0.85)	4.01 ±0.38 (3.39 - 6.54)	3.30 ±0.80 (2.70 - 13.50)	0.73 ±0.10 (0.09 - 0.84)	10.84 ±4.48 (5 - 25)
FREEZING TEMPERATURE SVDPLS	5.52 ±0.67 (3.80 - 8.92)	0.79 ±0.07 (0.40 - 0.94)	0.76 ±0.08 (0.16 - 0.84)	4.09 ±0.36 (3.51 - 6.23)	3.40 ±0.70 (2.70 - 11.00)	0.72 ±0.10 (0.07 - 0.83)	12.66 ±4.29 (5 - 25)
FREEZING TEMPERATURE PPLS	6.22 ±0.73 (4.25 - 10.05)	0.75 ±0.08 (0.16 - 0.90)	0.74 ±0.04 (0.53 - 0.83)	4.41 ±0.29 (3.71 - 5.42)	3.50 ±0.30 (2.90 - 5.30)	0.69 ±0.08 (0.30 - 0.82)	17.29 ±6.23 (6 - 25)
TOTAL AROMATICS MCSOR	1.07 ±0.32 (0.60 - 2.05)	0.98 ±0.02 (0.92 - 0.99)	0.91 ±0.08 (0.57 - 0.99)	0.86 ±0.14 (0.55 - 1.19)	1.10 ±0.50 (0.40 - 3.30)	0.91 ±0.08 (0.57 - 0.99)	14.56 ±6.23 (6 - 25)
TOTAL AROMATICS SIMPLS	1.59 ±0.72 (0.65 - 3.03)	0.94 ±0.05 (0.78 - 0.99)	0.88 ±0.08 (0.58 - 0.99)	1.14 ±0.39 (0.54 - 2.34)	1.40 ±0.60 (0.40 - 3.20)	0.87 ±0.08 (0.57 - 0.99)	14.65 ±7.13 (3 - 25)
TOTAL AROMATICS SVDPLS	1.02 ±0.26 (0.60 - 1.92)	0.98 ±0.01 (0.93 - 0.99)	0.93 ±0.07 (0.61 - 0.99)	0.82 ±0.13 (0.54 - 1.25)	1.00 ±0.50 (0.40 - 3.10)	0.93 ±0.07 (0.61 - 0.99)	15.21 ±5.48 (7 - 25)
TOTAL AROMATICS PPLS	1.23 ±0.21 (0.80 - 2.16)	0.97 ±0.01 (0.92 - 0.99)	0.97 ±0.02 (0.88 - 0.99)	0.84 ±0.12 (0.59 - 1.41)	0.70 ±0.10 (0.50 - 1.40)	0.96 ±0.02 (0.88 - 0.99)	23.35 ±2.57 (8 - 25)
VISCOSITY MCSOR	0.18 ±0.03 (0.11 - 0.29)	0.90 ±0.03 (0.73 - 0.95)	0.87 ±0.04 (0.70 - 0.94)	0.13 ±0.01 (0.10 - 0.18)	0.10 ±0.00 (0.12 - 0.23)	0.86 ±0.04 (0.68 - 0.94)	16.60 ±4.37 (6 - 25)
VISCOSITY SIMPLS	0.18 ±0.03 (0.11 - 0.27)	0.90 ±0.03 (0.75 - 0.95)	0.88 ±0.04 (0.72 - 0.93)	0.13 ±0.02 (0.10 - 0.19)	0.10 ±0.00 (0.15 - 0.21)	0.87 ±0.04 (0.64 - 0.93)	16.33 ±4.32 (6 - 25)
VISCOSITY SVDPLS	0.19 ±0.03 (0.12 - 0.31)	0.89 ±0.04 (0.73 - 0.95)	0.87 ±0.04 (0.70 - 0.93)	0.14 ±0.01 (0.11 - 0.19)	0.10 ±0.00 (0.10 - 0.20)	0.86 ±0.04 (0.70 - 0.93)	17.34 ±3.86 (8 - 25)
VISCOSITY PPLS	0.25 ±0.03 (0.16 - 0.34)	0.81 ±0.05 (0.59 - 0.92)	0.80 ±0.04 (0.65 - 0.90)	0.18 ±0.02 (0.13 - 0.24)	0.11 ±0.00 (0.13 - 0.22)	0.78 ±0.06 (0.56 - 0.90)	21.89 ±4.41 (5 - 25)

4.4.4 The performance of MCSOR in blind external tests

In general, thorough validation tests are essential for all modelling studies. However, the value of LOO CV as a model validation tool has been subject of much debate in the recent literature (see e.g. Golbraikh and Tropsha[401] vs. Hawkins et al. [511]). In practice, it is very easy to overfit a multivariate model and to get a biased and poorly predictive relationship with LOO CV. In particular, if the number of samples is large, LOO CV does not constitute a valid test for the predictive ability, as the estimated cross-validated correlation coefficient (Q^2) becomes too close to the conventional one, which is not of much value for underdetermined descriptor sets. In such cases, three data sets are recommended[61]. The first is the training set on which the cross-validation performance and the number of components is evaluated. This can be accomplished e.g. by LMO CV with a large number of randomised teaching and test sets, i.e. by choosing two-thirds of the training set compounds for the teaching set at random and placing the remaining in the test set. This inner loop is to be repeated several times (50 – 100, for example), and the validity is taken to be good if all pseudo-external performance indicators (R^2_{ex}, $|\Delta|_{ave}$, SDEP and Pr-r^2) are in the acceptable range and their scatter (standard deviations) is not very large. For MCSOR, the procedure also gives an estimate (on the basis of the Pr-r^2 maximum and its average, for example) for the number of components, i.e. the number of independent variables in the MLR models. The third data set is a truly blind prediction set which is not touched until the MCSOR model (i.e. master vectors, regression coefficients for predictors including the intercept, and the number of components) has been postulated completely. The whole training set is used to calculate the parameters. If wanted, this outer loop can be repeated at random several times.

In order to estimate the real performance of MCSOR, large QSAR/QSPR data sets with 2D MOE or DRAGON descriptors were used. In these cases, the number of samples is very large, so that they should provide a stringent test for the performance of MCSOR in a real situation i.e. in truly "blind" external predictions. The DRAGON descriptors were calculated directly from the SMILES codes so that they cannot account for any three-dimensional or charge effects (as neither coordinates nor partial charges of atoms are available); the descriptors that were either constant or zero were discarded, after that they were autoscaled. In general, the results will also indicate whether reasonable and computationally feasible QSAR/QSPR models can be derived for large and structurally diverse sets of organic molecules with MCSOR and restricted 1D/2D subsets of MOE or DRAGON descriptors.

The first data set, provided by Karthikeyan et al.[512], consists of 4173 organic molecules with their melting points and 2D/3D MOE descriptors. The 2D descriptors were selected for MCSOR modelling; the total number of them was 145. The second data set, provided by Fontaine et al.[513], contains a series of inhibitors of factor Xa, coming from very diverse chemical classes, but all sharing a benzamidine moiety. There are 435 compounds altogether, 156 of low-activity (K_i higher than 1μM) and 279 high-activity (K_i lower than 10 nM) compounds. After discarding the zero and constant DRAGON variables, the number of descriptors was 743. This data set, originally provided by Huuskonen et al.[514,515], consists of over 1300 organic compounds with logP and logS values. The DRAGON descriptors were calculated directly from the SMILES codes, but now all "molecular properties" descriptors were excluded from the model-

ling, as this category contains calculated log P and log S values which would have been too dominating in the QSPR models. After discarding the zero and constant variables, the number of descriptors was 829. All three data sets are freely available on the Web at http://cheminformatics.org/datasets/index.shtml.

In "blind" external tests, the performance of MCSOR varies considerably from case to case. Here again, we wish to emphasise that the primary aim of this study was not to develop alternatives for the original models, but to test the performance and validity of the MCSOR prediction algorithm. For the large melting point data set, the performance of MCSOR (Table 7) is only modest, being however slightly better (as assessed by root-mean square error, denoted as SDEP here) to that of the original models by Karthikeyan et al[512]. Nevertheless, the models leave much room for improvements, suggesting that this important physical property forms a challenging problem for QSPR methods. As with Karthikeyan et al.[512], it appeared that the incorporation of the 3D MOE descriptors did not improve the models. In general, a theoretical analysis of the melting point has been most elusive, as the almost complete lack of related publications clearly indicates[512]. For the factor Xa data set, the performance of MCSOR is quite satisfactory (Table 7), being again fully comparable to that of the original PLS model [513]. It should be emphasised that this data set is actually semi-quantitative, and thus the results indicate that MCSOR is feasible also for classification problems. For logP and logS data sets, the performance of MCSOR is surprisingly good (Table 7), bearing in mind the simplistic starting points of the model. In particular, the results indicate that the procedure suggested for the model validation, i.e. the use of inner and outer loops with three data sets, is reliable. Moreover, the success of the SMILES/DRAGON/MCSOR models demonstrates clearly that (i) 1D and 2D DRAGON descriptors contain much relevant information about hydrophobic properties and aqueous solubility of molecules, and (ii) MCSOR is able to extract useful information from a complex descriptor set. In comparison with the original, highly sophisticated models by Huuskonen et al. [514,515] that employ neural networks etc., the performance of MCSOR is actually only slightly worse. From a general point of view, a method that is successful for the prediction of log P values is expected to be feasible for many kinds of QSAR problems, as this physical quantity is overwhelmingly important for biological activity.

Table 7. Statistical performance indicators for blind MCSOR external predictions for melting point (MP), factor Xa, logP and log S predictions. Values are averages (± std) over 50 randomised runs where the sizes of teaching, test and "blind" prediction sets were N_{teach}, N_{test} and N_{ext} respectively.

Property	N_{teach}	N_{test}	N_{ext}		R^2_{ex}	$\|\Delta\|_{ave}$	SDEP	Pr-r^2	NPC
MP	1854	928	1391	**MCSOR**	0.49±0.02	35.1±0.56	46.1±0.78	0.49±0.02	32.2±6.15
				SIMPLS	0.49±0.02	35.8±0.58	46.2±0.85	0.49±0.02	31.6±4.59
				SVDPLS	0.40±0.03	39.6±0.82	50.2±1.11	0.39±0.28	4.26±0.53
				PPLS	0.39±0.02	40.0±1.02	50.8±1.18	0.38±0.02	17.6±4.86
FactorXa	193	97	145	**MCSOR**	0.79±0.02	1.51±0.10	2.00±0.12	0.79±0.03	8.51±1.81
				SIMPLS	0.79±0.02	1.57±0.12	2.04±0.15	0.78±0.03	9.06±2.85
				SVDPLS	0.71±0.06	1.85±0.14	2.39±0.24	0.69±0.07	3.88±0.33
				PPLS	0.76±0.04	1.69±0.16	2.20±0.19	0.74±0.05	15.1±2.66
log P	581	291	436	**MCSOR**	0.92±0.01	0.36±0.02	0.58±0.05	0.92±0.01	9.22±0.93
				SIMPLS	0.92±0.01	0.36±0.03	0.60±0.06	0.91±0.02	10.2±1.74
				SVDPLS	0.85±0.02	0.60±0.05	0.81±0.07	0.84±0.02	3.58±0.50
				PPLS	0.90±0.02	0.47±0.04	0.68±0.06	0.89±0.02	15.5±3.67
log S	579	290	435	**MCSOR**	0.90±0.01	0.51±0.02	0.67±0.02	0.89±0.01	9.34±0.66
				SIMPLS	0.89±0.01	0.52±0.03	0.68±0.03	0.89±0.01	9.14±0.75
				SVDPLS	0.81±0.01	0.70±0.03	0.90±0.04	0.81±0.02	3.02±0.14
				PPLS	0.86±0.01	0.60±0.03	0.78±0.03	0.85±0.01	14.9±0.76

4.5 The Pros and Cons of MCSOR

Conceptually, the MCSOR algorithm has many common features with other well-established multivariate methods. In fact, it can be viewed as a "hybrid" method that is composed of partial least-squares (PLS), principal components regression (PCR) and multiple linear regression (MLR). Yet, the MCSOR algorithm is among the simplest multivariate methods ever presented, both conceptually and computationally. It avoids all references to the eigenvalue theory of matrices while at the same time it provides results that are fully comparable to those of more advanced multivariate methods.

First and foremost, MCSOR shares many common characteristics with PLS. Like in all PLS algorithms, the dependent variable Y is used to derive the nearly orthogonal latent variables. In practice, small deviations of the master vectors from non-orthogonality seem not to be detrimental for successful predictions. As in PCR, the latent variables are used as independent variables to derive a MLR model between the Y-vector and the X-block vectors. The final prediction takes place via a simple regression equation, in which each term represents a nearly independent component. Nevertheless, it should be emphasised that the predictors are not usually orthogonal. Thus the MCSOR algorithm combines successfully some of the best features of the PLS, PCR and MLR algorithms. In particular, the master vectors are derived in a data-driven manner to ensure that the X- and Y-blocks will get information about each others. Thereby MCSOR avoids a common pitfall of PCR techniques by ensuring that the X-scores are strongly correlated with the Y vector. In fact, a method called continuum regression (CR) introduced by Stone and Brooks[516] has previously melded the MLR, PCR and PLS in a more formal way. In comparison with CR, MCSOR algorithm is more heuristic by nature.

In general, MCSOR has some advantages over PLS-type multivariate methods. A prominent feature of MCSOR is that it is not very sensitive to the number of components. This is due to the fact that the extra components have only a negligible contribution to the models, as the master vectors and predictors converge very rapidly towards zero. Thus the correct number of principal components to be selected, which is perhaps the most import problem with PLS, becomes less important as MCSOR seems to tolerate a large number of components without detrimental effects on external predictions. Second, MCSOR facilitates comparison between descriptors of different lengths, providing means to examine if different data sets really contain independent information for modelling and prediction. Third, the computational speed of SOR is high, facilitating the rapid development of model ensembles for consensus modelling[101,102]. In fact, the consensus approach could be applied in a straightforward way with MCSOR by taking an average over the individual MLR models with different number of components (which varies slightly in the inner validation loop) as a final prediction. A small number of components represents a situation where the fit is poor, but the predictive ability good and *vice versa*, so that an average over them should always represent a reasonable compromise.

The darker side of MCSOR is that its parameters are not particularly useful for explorative data analysis, in comparison e.g. with PCA and PLS scores and loadings. Of course, the plot of master vectors as a function of the variables reveals easily the most important variables for each MCSOR component. Moreover, in its current formulation the MCSOR is not applicable to mul-

tiple-Y problems, but each Y-vector must be handled separately. This is a clear disadvantage in cases where the Y-vectors are strongly correlated and thus they should be handled together. Nevertheless, these shortcomings are a cheap price to pay for simplicity and improved accuracy (in QSAR/QSPR problems with small data sets) for predictions.

Of course, the MCSOR algorithm also shares all conceptual and statistical ambiguities that are common for all "soft" modelling methods discussed recently by Helland[517] in the context of PLS. In particular, it should be emphasised that the correlations between X and Y derived by MCSOR are "made", not "natural" as is the case with PCR and standard MLR, for example. The present authors are inclined to think that this is actually one of the most dubious features of the data-driven modelling methods. Among other things, this makes the lateral verification[518] of the (QSAR) models difficult, if not impossible, as all the MCSOR parameters are derived from a particular data set, being valid and feasible for predictions in this limited context, but not transferable to any other data set. In this regard, all data-driven models contrast sharply e.g. with Hansch-type regression models, and their interpretability and transparency are thus seriously limited.

Finally, if there is a non-linear structure in the data, it is likely that non-linear modelling methods such as kNN (k nearest neighbour) or neural networks will outperform MCSOR and other linear methods. On the other hand, there are promising non-linear PLS extensions available such as spline-PLS[519], quadratic-PLS[520], and GIFI-PLS[521]. In principle, the corresponding extensions should be possible with MCSOR.

5. CONCLUSIONS AND FUTURE PROSPECTS

The results presented in this work clearly indicate that FLUFF-BALL is capable of generating robust prediction models for several different data sets and biological activities. In general the FLUFF-BALL generated results are comparable to those reported in literature. For highly congeneric systems BALL was slightly inferior, but on the other hand the diverse xenoestrogen data set BALL met or exceeded the results of the standard 3D-QSAR method CoMFA. The three variants of FLUFF superposition (FIX, MIX, and FLEX) allow rigid, semiflexible, and fully flexible superposition which efficiently leverages available *a priori* information in the form of user specified weight factors. The results also indicate that the FLUFF method is a versatile superposition technique which is suitable not only for BALL but also for other QSAR techniques such as CoMFA. Also, as the FLUFF-BALL technique is computationally simple and can easily be automated, it makes a useful and quite fast "molecular sieve". Despite this design emphasis, and the very low number descriptor variables used in BALL field, the FLUFF-BALL produced results comparable to the other QSAR techniques. Also one must bear in mind that the majority of them require an extensive amount of user involvement to produce the results and are therefore less suitable for fast screening applications.

The proof of equivalence between self-organizing regression (SOR) and SIMPLS with one principal component presented in section 4.1 clearly indicates the reason for SOMFA's poor performance when applied to complex datasets. Due to this limitation to one principal component the SOR, and by extension SOMFA, is suitable only for relatively simple cases where one principle component is sufficient. In a general case a more advanced regression method should be used instead. On the other hand, the results presented earlier clearly indicate that the applicability SOMFA can be successfully extended by replacing the SOR with a more complex regression method. Also, a novel multivariate regression method, *viz*. Multi Component Self-Organizing Regression (MCSOR), which is a multicomponent method loosely based upond the SOR philosophy, was presented. Extensive validation runs clearly indicate that the MCSOR is a promising alternative and supplement to more established multivariate methods. Therefore, it will be of considerable interest to test further its performance, besides PLS, against other soft modelling methods such as PCR, ridge regression (RR) and continuum regression (CR).

In general, it seems that in the 75 years since Hammett equation, the structure response correlation has had many failures, but also many successes. It is mainly due to the many success stories that the SRC has become an important tool for many disciplines where it has also found an important role as a complementary tool expanding and refining the information available for and from experimental work. Nevertheless, at the same time it must be emphasised that even though the SRC never lived up to be the infallible oracle of Delphoi as the most optimistic proponents hoped, and neither has it been as useless as the most pessimistic disponents feared. All in all the SRC in its myriad forms, despite its limitations, has proven to be an invaluable tool and therefore, one can with confidence predict that the SRC will play a more important role in the future. However, it is important that the practitioners of SRC remain cognisant of the limitations of these techniques and do not lose the critical eye of a sceptic. In particular, methodological breakthroughs are needed in order to increase the generalisability and stability of the analysis methods which at this time leave lot to be desired for. In part this can be also seen as a prob-

lem with the statistical analysis methodologies, but as was discussed in Section 2.5, a more powerful statistical technique does not necessarily mean a better model. Therefore a clear emphasis should be placed on the stringent validation of SRC models at all phases of their development. In my opinion one simply can not have a too stringent validation of SRC model and it is safer to err on the side of caution than to make overly optimistic interpretations of the results. Despite the cautionary examples presented in review articles[4,5,409,410] a horde of poorly validated models have been published which has been detrimental for the whole field of SRC. In order to minimise the chance of poor models, a considerable effort should also be devoted into finding a set of indicators which could be used to estimate the reliability of external predictions. In other words, the field of SRC would greatly benefit from a indicator, or set of indicators, which would warn the user when the reliability of the model is compromised.

One clearly positive aspect on the development of new SRC methodologies is the ever increasing computing capacity which undoubtedly will allow a great increase in the flexibility and power of the SRC techniques in the future. Despite this, the field of SRC is still far from maturity and it is very likely that in the coming 75 years the methodological advances will be as great as first 75 years of SRC. In my opinion one of the great challenges in developing the new generation of QSAR/QSPR methodologies lies in the incorporation of the dynamic nature of the molecules. Thus far the toy model of chemistry, static hard spheres like the plastic models, has dominated the field, but now it has been recognised that the molecules are dynamic and adaptable entities. The nD-QSAR methodologies proposed by Vedani et al[95,291,318,440] and the FLUFF-BALL[465,522] presented here are surely but the first attempts to encompass the true complexity of the biological and chemical systems. Therefore it seems clear that there is still much work to be done before the field of SRC analysis has reached maturity.

Reference List

1. Müller, G. nD QSAR: a Medicinal Chemist's point of view. *Quant. Struct. -Act. Relat.* **2002,** *21* (4), 391-396.
2. Kubinyi, H. From narcosis to hyperspace: The history of QSAR. *Quant. Struct. -Act. Relat.* **2002,** *21* (4), 348-356.
3. Combes, R.; Barratt, M.; Balls, M. An overall strategy for the testing of chemicals for human hazard and risk assessment under the EU REACH system. *ATLA* **2003,** *31* (1), 7-19.
4. Kim, K. H.; Greco, G.; Novellino, E. A Critical review of recent CoMFA applications. *Perspect. Drug Discov. Design* **1998,** *12/13/14*, 257-315.
5. Doweyko, A. M. 3D-QSAR illusions. *J. Comput. -Aided Mol. Des.* **2004,** *18* (7), 587-596.
6. Clark, D. E.; Pickett, S. D. Computational Methods for the Prediction of 'drug-likeness'. *Drug Discov. Today* **2000,** *5* (2), 49-58.
7. Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of methods for modelling quantitative structure-activity relationships. *J. Med. Chem.* **2004,** *47* (22), 5541-5554.
8. Rekker, R. F. The History of Drug Research - from Overton to Hansch. *Quant. Struct. -Act. Relat.* **1992,** *11* (2), 195-199.
9. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors;* Wiley-VCH: Weinheim, 2000.
10. Hansch, C.; Fujita, T. *p*–σ–π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964,** *86* (8), 1616-1626.
11. Beger, R. D.; Wilkes, J. G. Developing C-13 NMR quantitative spectrometric data-activity relationship (QSDAR) models of steroid binding to the corticosteroid binding globulin. *J. Comput. -Aided Mol. Des.* **2001,** *15* (7), 659-669.
12. Liu, Y.; Liu, S. S.; Cui, S. H.; Cai, S. X. A novel quantitative structure-biodegradability relationship (QSBR) of substituted benzenes based on MHDV descriptor. *J. Chin. Chem. Soc.* **2003,** *50* (2), 319-324.
13. Vandewaterbeemd, H. The History of Drug Research - from Hansch to the Present. *Quant. Struct. -Act. Relat.* **1992,** *11* (2), 200-204.
14. Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988,** *110* (18), 5959-5967.
15. Goodford, P. J. A Computational Procedure for Determining Energetically Favourable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985,** *28* (7), 849-857.
16. Kim, K. H. List of CoMFA references, 1993-1997. *Perspect. Drug Discov. Design* **1998,** *12*, 317-338.
17. Thibaut, U.; Folkers, G.; Klebe, G.; Kubinyi, H.; Merz, A.; Rognan, D. Recommendations for Comfa Studies and 3D Qsar Publications. *Quant. Struct. -Act. Relat.* **1994,** *13* (1), 1-3.
18. Melani, F.; Gratteri, P.; Adamo, M.; Bonaccini, C. Field interaction and geometrical overlap: A new simplex and experimental design based computational procedure for superposing small ligand molecules. *J. Med. Chem.* **2003,** *46* (8), 1359-1371.

19. Martin, Y. C. 3D QSAR: Current state, scope and limitations. *Perspect. Drug Discov. Design* **1998,** *12/13/14*, 3-23.
20. Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000,** *43* (17), 3233-3243.
21. Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput. -Aided Mol. Des.* **2000,** *14* (3), 215-232.
22. Ekins, S.; Waller, C. L.; Swaan, P. W.; Cruciani, G.; Wrighton, S. A.; Wikel, J. H. Progress in predicting human ADME parameters in silico. *J. Pharmacol. Toxicol. Methods* **2000,** *44* (1), 251-272.
23. Ekins, S.; Wrighton, S. A. Application of in silico approaches to predicting drug-drug interactions. *J. Pharmacol. Toxicol. Methods* **2001,** *45* (1), 65-69.
24. Liu, R. F.; So, S. S. Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 1. Aqueous solubility. *J. Chem. Inf. Comp. Sci.* **2001,** *41* (6), 1633-1639.
25. Liu, R. F.; Sun, H. M.; So, S. S. Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 2. Blood-brain barrier penetration. *J. Chem. Inf. Comp. Sci.* **2001,** *41* (6), 1623-1632.
26. Wessel, M. D.; Mente, S. Chapter 25. ADME by computer. *Annu. Rep. Med. Chem.* **2001,** *36*, 257-266.
27. Balaz, S.; Lukacova, V. Subcellular pharmacokinetics and its potential for library focusing. *J. Mol. Graph. Model.* **2002,** *20* (6), 479-490.
28. Butina, D.; Segall, M. D.; Frankcombe, K. Predicting ADME properties in silico: methods and models. *Drug Discov. Today* **2002,** *7* (11), S83-S88.
29. Egan, W. J.; Walters, W. P.; Murcko, M. A. Guiding molecules towards drug-likeness. *Curr. Opin. Drug Discovery Dev.* **2002,** *5* (4), 540-549.
30. Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery - 1. Applications of genetic algorithms to the prediction of blood-brain partitioning of a large set of drugs. *J. Mol. Model.* **2002,** *8* (12), 337-349.
31. Keseru, G. M.; Molnar, L. METAPRINT: A metabolic fingerprint. Application to cassette design for high-throughput ADME screening. *J. Chem. Inf. Comp. Sci.* **2002,** *42* (2), 437-444.
32. Klein, C.; Kaiser, D.; Kopp, S.; Chiba, P.; Ecker, G. F. Similarity based SAR (SIBAR) as tool for early ADME profiling. *J. Comput. -Aided Mol. Des.* **2002,** *16* (11), 785-793.
33. Blaauboer, B. J. The integration of data on physico-chemical properties, in vitro-derived toxicity data and physiologically based kinetic and dynamic as modelling a tool in hazard and risk assessment. A commentary. *Toxicol. Lett.* **2003,** *138* (1-2), 161-171.
34. Colmenarejo, G. In silico prediction of drug-binding strengths to human serum albumin. *Med. Res. Rev.* **2003,** *23* (3), 275-301.
35. Hou, T. J.; Xu, X. J. ADME evaluation in drug discovery. 3. Modeling blood-brain barrier partitioning using simple molecular descriptors. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (6), 2137-2152.
36. Hutter, M. C. Prediction of blood-brain barrier permeation using quantum chemically derived information. *J. Comput. -Aided Mol. Des.* **2003,** *17* (7), 415-433.
37. Shen, M.; Xiao, Y. D.; Golbraikh, A.; Gombar, V. K.; Tropsha, A. Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J. Med. Chem.* **2003,** *46* (14), 3013-3020.

38. Theil, F. P.; Guentert, T. W.; Haddad, S.; Poulin, P. Utility of physiologically based pharmacokinetic models to drug development and rational drug discovery candidate selection. *Toxicol. Lett.* **2003,** *138* (1-2), 29-49.
39. Wolohan, P. R. N.; Clark, R. D. Predicting drug pharmacokinetic properties using molecular interaction fields and SIMCA. *J. Comput. -Aided Mol. Des.* **2003,** *17* (1), 65-76.
40. Chen, H. F.; Yao, X. J.; Petitjean, M.; Xia, H. O.; Yao, J. H.; Panaye, A.; Doucet, J. P.; Fan, B. T. Insight into the bioactivity and metabolism of human glucagon receptor antagonists from 3D-QSAR analyses. *QSAR Comb. Sci.* **2004,** *23* (8), 603-620.
41. Crivori, P.; Zamora, I.; Speed, B.; Orrenius, C.; Poggesi, I. Model based on GRID-derived descriptors for estimating CYP3A4 enzyme stability of potential drug candidates. *J. Comput. -Aided Mol. Des.* **2004,** *18* (3), 155-166.
42. Ekins, S.; Swaan, P. W. Development of computational models for enzymes, transporters, channels, and receptors relevant to ADME/Tox. *Rev. Comp. Chem.* **2004,** *20*, 333-415.
43. Eros, D.; Keri, G.; Kovesdi, I.; Szantai-Kis, C.; Meszaros, G.; Orfi, L. Comparison of predictive ability of water solubility QSPR models generated by MLR, PLS and ANN methods. *Mini-Rev. Med. Chem.* **2004,** *4* (2), 167-177.
44. Hansch, C.; Mekapati, S. B.; Kurup, A.; Verma, R. P. QSAR of cytochrome P450. *Drug Metab. Rev.* **2004,** *36* (1), 105-156.
45. Hansch, C.; Leo, A.; Mekapati, S. B.; Kurup, A. Qsar and Adme. *Biorg. Med. Chem.* **2004,** *12* (12), 3391-3400.
46. Hemmateenejad, B. Optimal QSAR analysis of the carcinogenic activity of drugs by correlation ranking and genetic algorithm-based. *J. Chemom.* **2004,** *18* (11), 475-485.
47. Hou, T. J.; Zhang, W.; Xia, K.; Qiao, X. B.; Xu, X. J. ADME evaluation in drug discovery. 5. Correlation of Caco-2 permeation with simple molecular properties. *J. Chem. Inf. Comp. Sci.* **2004,** *44* (5), 1585-1600.
48. Nordqvist, A.; Nilsson, J.; Lindmark, T.; Eriksson, A.; Garberg, P.; Kihlen, M. A general model for prediction of Caco-2 cell permeability. *QSAR Comb. Sci.* **2004,** *23* (5), 303-310.
49. Smith, P. A.; Sorich, M. J.; Low, L. S. C.; McKinnon, R. A.; Miners, J. O. Towards integrated ADME prediction: past, present and future directions for modelling metabolism by UDP-glucuronosyltransferases. *J. Mol. Graph. Model.* **2004,** *22* (6), 507-517.
50. Stoner, C. L.; Gifford, E.; Stankovic, C.; Lepsy, C. S.; Brodfuehrer, J.; Prasad, J. V. N. V.; Surendran, N. Implementation of an ADME enabling selection and visualization tool for drug discovery. *J. Pharm. Sci.* **2004,** *93* (5), 1131-1141.
51. Weaver, D. C. Applying data mining techniques to library design, lead generation and lead optimization. *Curr. Opin. Chem. Biol.* **2004,** *8* (3), 264-270.
52. Winkler, D. A. Neural networks in ADME and toxicity prediction. *Drugs Fut.* **2004,** *29* (10), 1043-1057.
53. Winkler, D. A.; Burden, F. R. Modelling blood-brain barrier partitioning using Bayesian neural nets. *J. Mol. Graph. Model.* **2004,** *22* (6), 499-505.
54. Andricopulo, A. D.; Montanari, C. A. Structure-activity relationships for the design of small-molecule inhibitors. *Mini-Rev. Med. Chem.* **2005,** *5* (6), 585-593.

55. Hemmateenejad, B.; Safarpour, M. A.; Miri, R.; Nesari, N. Toward an optimal procedure for PC-ANN model building: Prediction of the carcinogenic activity of a large set of drugs. *J. Chem. Inf. Mod.* **2005,** *45* (1), 190-199.
56. Narayanan, R.; Gunturi, S. B. In silico ADME modelling: prediction models for blood-brain barrier permeation using a systematic variable selection method. *Biorg. Med. Chem.* **2005,** *13* (8), 3017-3028.
57. Votano, J. R. Recent uses of topological indices in the development of in silico ADMET models. *Curr. Opin. Drug Discovery Dev.* **2005,** *8* (1), 32-37.
58. Yap, C. W.; Chen, Y. Z. Quantitative structure-pharmacokinetic relationships for drug distribution properties by using general regression neural network. *J. Pharm. Sci.* **2005,** *94* (1), 153-168.
59. Zhao, S. W.; Liu, L.; Fu, Y.; Guo, Q. X. Assessment of the metabolic stability of the methyl groups in heterocyclic compounds using C-H bond dissociation energies: effects of diverse aromatic groups on the stability of methyl radicals. *J. Phys. Org. Chem.* **2005,** *18* (4), 353-367.
60. Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Chen, X. Q.; Doweyko, A.; Li, Y. In silico ADME/Tox: why models fail. *J. Comput. -Aided Mol. Des.* **2003,** *17* (2), 83-92.
61. Norinder, U. In silico modelling of ADMET - A minireview of work from 2000 to 2004. *SAR QSAR Environ. Res.* **2005,** *16* (1-2), 1-11.
62. Melani, F.; Gratteri, P.; Adamo, M.; Bonaccini, C. FILO (Field Interaction Ligand Optimization): A simplex strategy for searching the optimal ligand interaction field in drug design. *J. Comput. -Aided Mol. Des.* **2001,** *15* , 57-66.
63. Sippl, W. Receptor-based 3D QSAR analysis of estrogen receptor ligands - merging the accuracy of receptor-based aligment with computational efficiency of ligand-based methods. *J. Comput. -Aided Mol. Des.* **2000,** *14* (6), 559-572.
64. Gratteri, P.; Bonaccini, C.; Melani, F. Searching for a reliable orientation of ligands in their binding site: Comparison between a structure-based (Glide) and a ligand-based (FIGO) approach in the case study of PDE4 inhibitors. *J. Med. Chem.* **2005,** *48* (5), 1657-1665.
65. Oprea, T. I.; Marshall, G. R. Receptor-based prediction of binding affinities. *Perspect. Drug Discov. Design* **1998,** *9/10/11*, 35-61.
66. Mestres, J.; Knegtel, R. M. A. Similarity versus docking in 3D virtual screening. *Perspect. Drug Discov. Design* **2000,** *20* (1), 191-207.
67. Krovat, E. M.; Steindl, T.; Langer, T. Recent Advances in Docking and Scoring. *Curr. Comput. -Aided Drug Des.* **2005,** *1*, 93-102.
68. Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A Review of Protein-small molecule docking methods. *J. Comput. -Aided Mol. Des.* **2002,** *16* (6), 151-166.
69. De Rosa, M. C.; Berglund, A. A New Method for Predicting the Alignment of Flexible Molecules and Orienting Them in a Receptor Cleft of Know Structure. *J. Med. Chem.* **1998,** *41* (5), 691-698.
70. Hare, B. J.; Walters, W. P.; Caron, P. R.; Bemis, G. W. CORES: An Automated method for generating three-dimensional models of protein/ligand complexes. *J. Med. Chem.* **2004,** *47* (19), 4731-4740.
71. Miller, M. D.; Sheridan, R. P.; Kearsley, S. K. SQ: A Program for Repidly Producing Pharmacophorically Relevant Molecular Superpositions. *J. Med. Chem.* **1999,** *42* (9), 1505-1514.

72. Arakawa, M.; Hasegawa, K.; Funatsu, K. Novel alignment method of small molecules using the Hopfield neural networks. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (5), 1390-1395.
73. Arakawa, M.; Hasegawa, K.; Funatsu, K. Application of novel alignment method of small molecules using the Hopfield neural networks to 3D-QSAR. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (5), 1396-1402.
74. Perkins, T. D. J.; Mills, J. E. J.; Dean, P. M. Molecular surface-volume and property matching to superpose flexible dissimilar molecules. *J. Comput. -Aided Mol. Des.* **1995,** *9* , 479-490.
75. Putta, S.; Eksterowicz, J.; Lemmen, C.; Stanton, R. A Novel subshape molecular descriptor. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (5), 1623-1635.
76. Hofbauer, C.; Lohninger, H.; Aszódi, A. SURFCOMP: A Novel Graph-Based Approach to Molecular Surface Comparison. *J. Chem. Inf. Comp. Sci.* **2004,** *44*, 837-847.
77. Diller, D. J.; Verlinde, C. L. M. J. A critical evaluation of several global optimization algorithms for the purpose of molecular docking. *J. Comp. Chem.* **1999,** *20* (16), 1740-1751.
78. McMartin, C.; Bohacek, R. QXP: Powerful, rapid computer algorithms for structure-based drug design. *J. Comput. -Aided Mol. Des.* **1997,** *11* (4), 333-344.
79. Kearsley, S. K.; Smith, G. M. An Alternative Method for the Alignment of Molecular Structures: Maximizing Electrostatic and Steric Overlap. *Tetrahedron Comput. Methodol.* **1990,** *3* (6C), 613-633.
80. McMartin, C.; Bohacek, R. Flexible matching of test ligands to a 3D pharmacophore using a molecular superposition force field: Comparison of Predicted and experimental conformations of inhibitors of three enzymes. *J. Comput. -Aided Mol. Des.* **1995,** *9* , 237-250.
81. Mestres, J.; Rohrer, D. C.; Maggiora, G. M. MIMIC: A Molecular-Field Matching Program Exploting Applicability of Molecular Similarity Approaches. *J. Comp. Chem.* **1997,** *18* (7), 934-954.
82. Thorner, D. A.; Wild, D. J.; Willett, P.; Wright, P. M. Calculation of structural similarity by the alignment of molecular electorstatic potentials. *Perspect. Drug Discov. Design* **1998,** *9/10/11*, 301-320.
83. Klebe, G.; Mietzner, T.; Weber, F. Methodological developments and strategies for a fast flexible superposition of drug-sized molecules. *J. Comput. -Aided Mol. Des.* **1999,** *13* (1), 35-49.
84. Krämer, A.; Horn, H. W.; Rice, J. E. Fast 3D molecular superposition and similarity search in databases of flexible molecules. *J. Comput. -Aided Mol. Des.* **2003,** *17* (1), 13-38.
85. Bultinck, P.; Kuppens, T.; Gironés, X.; Carbó-Dorca, R. Quantum Similarity Superposition Algorithm (QSSA): A Consistent Scheme for Molecular Alignment and Molecular Similarity Based on Quantum Chemistry. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (4), 1143-1150.
86. Robinson, D. D.; Lyne, P. D.; Richards, W. G. Partial Molecular Alignment via Local Structure Analysis. *J. Chem. Inf. Comp. Sci.* **2000,** *40* (2), 503-512.
87. Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. -Aided Mol. Des.* **2002,** *16* (7), 521-533.

88. Mills, J. E. J.; de Esch, I. J. P.; Perkins, T. D. J.; Dean, P. M. SLATE: A method for the superposition of flexible ligands. *J. Comput. -Aided Mol. Des.* **2001,** *15* , 81-96.
89. Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem.* **1998,** *41* (23), 4502-4520.
90. Berglund, A.; De Rosa, M. C.; Wold, S. Aligment of flexible molecules at their receptor site using 3D descriptors and Hi-PCA. *J. Comput. -Aided Mol. Des.* **1997,** *11* (6), 601-612.
91. Lemmen, C.; Zimmermann, M.; Lengauer, T. Multiple molecular superpositioning as an effective tool for virtual screening. *Perspect. Drug Discov. Design* **2000,** *20* (1), 43-62.
92. Good, A. C.; Richards, W. G. Explicit calculation of 3D molecular similarity. *Perspect. Drug Discov. Design* **1998,** *9/10/11*, 321-338.
93. Gironés, X.; Carbó-Dorca, R. TGSA-Flex: extending the capabilities of the topogeometrical superposition algorithm to handle flexible molecules. *J. Comp. Chem.* **2004,** *25* (2), 153-159.
94. *SYBYL. Computational Informatics Software for Molecular Modelers,* Tripos Inc.: 2006.
95. Vedani, A.; Dobler, M. 5D-QSAR: The key for simulating induced fit? *J. Med. Chem.* **2002,** *45* (11), 2139-2149.
96. *DRAGON. Software for calculation of molecular descriptors,* 2006.
97. *Codessa. Quantitative Structure/Activity Relationship (QSAR) program,* 2006.
98. Vedrina, M.; Markovic, S.; Medic-Saric, M.; Trinajstic, N. TAM: A program for the calculation of topological indices in QSPR and QSAR studies. *Comput. Chem.* **1997,** *21* (6), 355-361.
99. das Neves, P. J.; da Costa, J. B. N.; Ndiyae, P. M.; Carneiro, J. W. D. TOP - A software for calculation of topological descriptors to be used in structure-activity relationships. *Quimica Nova* **1998,** *21* (6), 709-713.
100. Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Chen, X.-Q.; Doweyko, A.; Li, Y. In silico ADME/TOX: why models fail. *J. Comput. -Aided Mol. Des.* **2003,** *17* (2), 83-92.
101. Asikainen, A. H.; Ruuskanen, J.; Tuppurainen, K. Consensus kNN QSAR: A Versatile Method for Predicting the Estrogenic Activity of Organic Compounds In Silico. A Comparative Study with Five Estrogen Receptors and a Large, Diverse Set of Ligands. *Environ. Sci. Technol.* **2004,** *38* (24), 6730-6740.
102. Asikainen, A. H.; Ruuskanen, J.; Tuppurainen, K. Performance of (consensus) kNN QSAR for predicting estrogenic activity in large diverse set of organic compounds. *SAR QSAR Environ. Res.* **2004,** *15* (1), 19-32.
103. de Julian-Ortiz, J. V.; Besalu, E.; Garcia-Domenech, R. True prediction by consensus for small sets of cyclooxigenase-2 inhibitors. *Ind. J. Chem.* **2003,** *42* (6), 1392-1404.
104. Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q. A.; Tong, W. Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis* **2004,** *19* (5), 365-377.
105. Beger, R. D.; Wilkes, J. G. Models of polychlorinated dibenzodioxins, dibenzofurans, and biphenyls binding affinity to the aryl hydrocarbon receptor developed using C-13 NMR data. *J. Chem. Inf. Comp. Sci.* **2001,** *41* (5), 1322-1329.

106. Beger, R. D.; Buzatu, D. A.; Wilkes, J. G.; Lay, J. O. C-13 NMR quantitative spectrometric data-activity relationship (QSDAR) models of steroids binding the aromatase enzyme. *J. Chem. Inf. Comp. Sci.* **2001,** *41* (5), 1360-1366.
107. Bursi, R.; Dao, T.; van Wijk, T.; de Gooyer, M.; Kellenbach, E.; Verwer, P. Comparative spectra analysis (CoSA): Spectra as three-dimensional molecular descriptors for the prediction of biological activities. *J. Chem. Inf. Comp. Sci.* **1999,** *39* (5), 861-867.
108. Beger, R. D.; Buzatu, D. A.; Wilkes, J. G. Combining NMR spectral and structural data to form models of polychlorinated dibenzodioxins, dibenzofurans, and biphenyls binding to the AhR. *J. Comput. -Aided Mol. Des.* **2002,** *16* (10), 727-740.
109. Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Sheehan, D. M. Evaluation of quantitative structure-activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. *J. Chem. Inf. Comp. Sci.* **1998,** *38* (4), 669-677.
110. So, S. S.; Karplus, M. A comparative study of ligand-receptor complex binding affinity prediction methods based on glycogen phosphorylase inhibitors. *J. Comput. -Aided Mol. Des.* **1999,** *13* (3), 243-258.
111. Cho, S. G.; Goh, E. M.; Kim, J. K. Holographic QSAR models for estimating densities of energetic materials. *Bull. Kor. Chem. Soc.* **2001,** *22* (7), 775-778.
112. Choo, H. Y. P.; Lim, J. S.; Kam, Y.; Kim, S. Y.; Lee, J. A comparative study of quantitative structure activity relationship methods based on antitumor diarylsulfonylureas. *Eur. J. Med. Chem.* **2001,** *36* (10), 829-836.
113. Huang, X. Q.; Liu, T.; Gu, J. D.; Luo, X. M.; Ji, R. Y.; Cao, Y.; Xue, H.; Wong, J. T. F.; Wong, B. L.; Pei, G.; Jiang, H. L.; Chen, K. X. 3D-QSAR model of flavonoids binding at benzodiazepine site in GABA(A) receptors. *J. Med. Chem.* **2001,** *44* (12), 1883-1891.
114. Rodrigues, C. R.; Flaherty, T. M.; Springer, C.; McKerrow, J. H.; Cohen, F. E. CoMFA and HQSAR of acylhydrazide cruzain inhibitors. *Biorg. Med. Chem. Lett.* **2002,** *12* (11), 1537-1541.
115. Suh, M. E.; Park, S. Y.; Lee, H. J. Comparison of QSAR methods (CoMFA, CoMSIA, HQSAR) of anticancer 1-N-substituted imidazoquinoiine-4,9-dione derivatives. *Bull. Kor. Chem. Soc.* **2002,** *23* (3), 417-422.
116. Cui, S. H.; Liu, S. S.; Wang, X. D.; Wang, L. S. Holographic QSAR of estradiol derivatives. *Chin. Sci. Bull.* **2003,** *48* (7), 642-645.
117. Chen, D.; Yin, C. S.; Wang, X. D.; Wang, L. S. Holographic QSAR of selected esters. *Chemosphere* **2004,** *57* (11), 1739-1745.
118. Chen, H. F.; Li, Q.; Yao, X. J.; Fan, B. T.; Yuan, S. G.; Panaye, A.; Doucet, J. P. CoMFA/CoMSIA/HQSAR and docking study of the binding mode of selective cyclooxygenase (COX-2) inhibitors. *QSAR Comb. Sci.* **2004,** *23* (1), 36-55.
119. Huang, H.; Wang, X. D.; Dai, X. L.; Yu, Y. J.; Wang, L. S. Holographic quantitative structure-activity relationship for prediction acute toxicity of benzene derivatives to the guppy(poecilia reticulata). *J. Environ. Sci.* **2004,** *16* (3), 423-427.
120. Huang, H.; Ou, W.; Zhao, J.; Chen, D.; Wang, L. A comparative study of quantitative structure-activity relationship methods based on gallic acid derivatives. *SAR QSAR Environ. Res.* **2004,** *15* (2), 83-99.

121. Song, Y. S.; Sung, N. D.; Yu, Y. M.; Kim, B. T. QSAR studies on the inhibitory activity of new methoxyacrylate analogues against Magnaporthe grisea (Rice Blast Disease). *Bull. Kor. Chem. Soc.* **2004,** *25* (10), 1513-1520.
122. Waller, C. L. A comparative QSAR study using CoMFA, HQSAR, and FRED/SKEYS paradigms for estrogen receptor binding affinities of structurally diverse compounds. *J. Chem. Inf. Comp. Sci.* **2004,** *44* (2), 758-765.
123. Cho, S. J. Hologram quantitative structure activity relationship (HQSAR) study of mutagen X. *Bull. Kor. Chem. Soc.* **2005,** *26* (1), 85-90.
124. Cunningham, S. L.; Cunningham, A. R.; Day, B. W. CoMFA, HQSAR and molecular docking studies of butitaxel analogues with beta-tubulin. *J. Mol. Model.* **2005,** *11* (1), 48-54.
125. Honorio, K. M.; Garratt, R. C.; Andricopulo, A. D. Hologram quantitative structure-activity relationships for a series of farnesoid X receptor activators. *Biorg. Med. Chem. Lett.* **2005,** *15* (12), 3119-3125.
126. Zhang, H. B.; Li, H.; Liu, C. P. CoMFA, CoMSIA, and molecular hologram QSAR studies of novel neuronal nAChRs ligands-open ring analogues of 3-pyridyl ether. *J. Chem. Inf. Mod.* **2005,** *45* (2), 440-448.
127. Zhu, W. L.; Gang, C.; Hu, L. H.; Luo, X. M.; Gui, C. S.; Cheng, L.; Puah, C. M.; Chen, K. X.; Jiang, H. L. QSAR analyses on ginkgolides and their analogues using CoMFA, CoMSIA, and HQSAR. *Biorg. Med. Chem.* **2005,** *13* (2), 313-322.
128. Castro, R. A.; Gutman, I.; Marino, D.; Peruzzo, P. Upgrading the Wiener index. *J. Serb. Chem. Soc.* **2002,** *67* (10), 647-651.
129. Diudea, M. V.; Gutman, I. Wiener-type topological indices. *Croat. Chem. Acta* **1998,** *71* (1), 21-51.
130. Garcia, G. C.; Ruiz, I. L.; Gomez-Nieto, M. A. From Wiener index to molecules. *J. Chem. Inf. Mod.* **2005,** *45* (2), 231-238.
131. Gutman, I.; Linert, W.; Lukovits, I.; Tomovic, Z. The multiplicative version of the Wiener index. *J. Chem. Inf. Comp. Sci.* **2000,** *40* (1), 113-116.
132. Gutman, I.; Vukicevic, D.; Zerovnik, J. A class of modified Wiener indices. *Croat. Chem. Acta* **2004,** *77* (1-2), 103-109.
133. Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D.; Balaban, A. T. Comparison of weighting schemes for molecular graph descriptors: Application in quantitative structure - Retention relationship models for alkylphenols in gas-liquid chromatography. *J. Chem. Inf. Comp. Sci.* **2000,** *40* (3), 732-743.
134. Ivanciuc, O.; Ivanciuc, T.; Klein, D. J.; Seitz, W. A.; Balaban, A. T. Wiener index extension by counting even/odd graph distances. *J. Chem. Inf. Comp. Sci.* **2001,** *41* (3), 536-549.
135. Li, X. H. The extended Wiener index. *Chem. Phys. Lett.* **2002,** *365* (1-2), 135-139.
136. Li, X. H.; Li, Z. G.; Hu, M. L. A novel set of Wiener indices. *J. Mol. Graph. Model.* **2003,** *22* (2), 161-172.
137. Li, X. H. The extended hyper-Wiener index. *Can. J. Chem.* **2003,** *81* (9), 992-996.
138. Randic, M. Search for Optimal Molecular Descriptors. *Croat. Chem. Acta* **1991,** *64* (1), 43-54.
139. Randic, M. Novel Molecular Descriptor for Structure-Property Studies. *Chem. Phys. Lett.* **1993,** *211* (4-5), 478-483.
140. Randic, M.; Guo, X. F.; Oxley, T.; Krishnapriyan, H.; Naylor, L. Wiener Matrix Invariants. *J. Chem. Inf. Comp. Sci.* **1994,** *34* (2), 361-367.

141. Sardana, S.; Madan, A. K. Application of graph theory: Relationship of molecular connectivity index, Wiener's index and eccentric connectivity index with diuretic activity. *MATCH-Commun. Math. Ch.* **2001,** (43), 85-98.
142. Bajaj, S.; Sambi, S. S.; Madan, A. K. Topological models for prediction of anti-inflammatory activity of N-arylanthranilic acids. *Biorg. Med. Chem.* **2004,** *12* (13), 3695-3701.
143. Bajaj, S.; Sambi, S. S.; Madan, A. K. Prediction of anti-inflammatory activity of N-arylanthranilic acids: Computational approach using refined Zagreb indices. *Croat. Chem. Acta* **2005,** *78* (2), 165-174.
144. Bonchev, D. Overall connectivity - a next generation molecular connectivity. *J. Mol. Graph. Model.* **2001,** *20* (1), 65-75.
145. Milicevic, A.; Nikolic, S. On variable Zagreb indices. *Croat. Chem. Acta* **2004,** *77* (1-2), 97-101.
146. Nikolic, S.; Kovacevic, G.; Milicevic, A.; Trinajstic, N. The Zagreb indices 30 years after. *Croat. Chem. Acta* **2003,** *76* (2), 113-124.
147. Delgado, E. J.; Matamala, A.; Alderete, J. B. Predicting gas chromatographic retention time of polychlorinated dibenzo-p-dioxins from molecular structure. *Z. Phys. Chem.* **2002,** *216* (4), 451-457.
148. Jalali-Heravi, M.; Konouz, E. Prediction of critical micelle concentration of some anionic surfactants using multiple regression techniques: A quantitative structure-activity relationship study. *J. Surfactants Deterg.* **2000,** *3* (1), 47-52.
149. Katritzky, A. R.; Gordeeva, E. V. Traditional Topological Induces Vs Electronic, Geometrical, and Combined Molecular Descriptors in Qsar Qspr Research. *J. Chem. Inf. Comp. Sci.* **1993,** *33* (6), 835-857.
150. Lucic, B.; Basic, I.; Nadramija, D.; Milicevic, A.; Trinajstic, N.; Suzuki, T.; Petrukhin, R.; Karelson, M.; Katritzky, A. R. Correlation of liquid viscosity with molecular structure for organic compounds using different variable selection methods. *Arkivoc* **2002,** 45-59.
151. Taherpour, A.; Shafiei, F. The structural relationship between Randic indices, adjacency matrixes, distance matrixes and maximum wave length of linear simple conjugated polyene compounds. *Theochem. J. Mol. Struct.* **2005,** *726* (1-3), 183-188.
152. Estrada, E.; Gutierrez, Y. The Balaban J index in the multidimensional space of generalized topological indices. Generalizations and QSPR improvements. *MATCH-Commun. Math. Ch.* **2001,** (44), 155-167.
153. Gupta, S.; Singh, A.; Madan, A. K. Connective eccentricity index: A novel topological descriptor for predicting biological activity. *J. Mol. Graph. Model.* **2000,** *18* (1), 18-25.
154. Nikolic, S.; Plavsic, D.; Trinajstic, N. On the Balaban-like topological indices. *MATCH-Commun. Math. Ch.* **2001,** (44), 361-386.
155. Osmialowski, K.; Kaliszan, R. Studies of Performance of Graph Theoretical Indexes in Qsar Analysis. *Quant. Struct. -Act. Relat.* **1991,** *10* (2), 125-134.
156. Roy, K.; Ghosh, G. QSTR with extended topochemical atom indices. 3. Toxicity of nitrobenzenes to Tetrahymena pyriformis. *QSAR Comb. Sci.* **2004,** *23* (2-3), 99-108.
157. Rucker, G.; Rucker, C. Counts of All Walks As Atomic and Molecular Descriptors. *J. Chem. Inf. Comp. Sci.* **1993,** *33* (5), 683-695.

158. Thakur, A.; Thakur, M.; Khadikar, P. V.; Supuran, C. T.; Sudele, P. QSAR study on benzenesulphonamide carbonic anhydrase inhibitors: topological approach using Balaban index. *Biorg. Med. Chem.* **2004,** *12* (4), 789-793.
159. Ivanciuc, O.; Ivanciuc, T.; Klein, D. J. Quantitative structure-property relationships generated with optimizable even/odd Wiener polynomial descriptors. *SAR QSAR Environ. Res.* **2001,** *12* (1-2), 1-+.
160. Jalbout, A. F.; Li, X. H. Bond order weighted Wiener numbers. *Theochem. J. Mol. Struct.* **2003,** *663* (1-3), 9-14.
161. Juvan, M.; Mohar, B. Bond Contributions to the Wiener Index. *J. Chem. Inf. Comp. Sci.* **1995,** *35* (2), 217-219.
162. Lim, T. C. Mass-modified Wiener indices and the boiling points for lower chloroalkanes. *Acta Chim. Slov.* **2004,** *51* (4), 611-618.
163. Butina, D. Performance of Kier-hall E-state descriptors in quantitative structure activity relationship (QSAR) studies of multifunctional molecules. *Molecules* **2004,** *9* (12), 1004-1009.
164. Lucic, B.; Nikolic, S.; Trinajstic, N.; Juric, A.; Mihalic, Z. A Structure-Property Study of the Solubility of Aliphatic-Alcohols in Water. *Croat. Chem. Acta* **1995,** *68* (3), 417-434.
165. Ren, B. Y. Novel atom-type AI indices for QSPR studies of alcohols. *Comput. Chem.* **2002,** *26* (3), 223-235.
166. Ren, B. Y. Novel atomic-level-based AI topological descriptors: Application to QSPR/QSAR modeling. *J. Chem. Inf. Comp. Sci.* **2002,** *42* (4), 858-868.
167. Ren, B. Y. Application of novel atom-type AI topological indices in the structure-property correlations. *Theochem. J. Mol. Struct.* **2002,** *586*, 137-148.
168. TsantiliKakoulidou, A.; Kier, L. B.; Joshi, N. The Use of Electrotopological State Indexes in Qsar Studies. *J. Chim. Phys. Phys. - Chim. Biol.* **1992,** *89* (7-8), 1729-1733.
169. Hall, L. H.; Kier, L. B. Electrotopological State Indexes for Atom Types - A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comp. Sci.* **1995,** *35* (6), 1039-1045.
170. AbouShaaban, R. R. A.; AlKhamees, H. A.; AbouAuda, H. S.; Simonelli, A. P. Atom level electrotopological state indexes in QSAR: Designing and testing antithyroid agents. *Pharm. Res.* **1996,** *13* (1), 129-136.
171. de Gregorio, C.; Kier, L. B.; Hall, L. H. QSAR modeling with the electrotopological state indices: Corticosteroids. *J. Comput. -Aided Mol. Des.* **1998,** *12* (6), 557-561.
172. Kier, L. B.; Hall, L. H. The E-state in database analysis: the PCBs as an example. *Farmaco* **1999,** *54* (6), 346-353.
173. Maw, H. H.; Hall, L. H. E-state modeling of HIV-1 protease inhibitor binding independent of 3D information. *J. Chem. Inf. Comp. Sci.* **2002,** *42* (2), 290-298.
174. Rose, K.; Hall, L. H.; Kier, L. B. Modeling blood-brain barrier partitioning using the electrotopological state. *J. Chem. Inf. Comp. Sci.* **2002,** *42* (3), 651-666.
175. Contrera, J. F.; Matthews, E. J.; Benz, R. D. Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Reg. Toxicol. Pharmacol.* **2003,** *38* (3), 243-259.
176. Debnath, B.; Samanta, S.; Naskar, S. K.; Roy, K.; Jha, T. QSAR study on the affinity of some arylpiperazines towards the 5-HT1A/alpha(1)-adrenergic receptor using the E-state index. *Biorg. Med. Chem. Lett.* **2003,** *13* (17), 2837-2842.

177. Hall, L. M.; Hall, L. H.; Kier, L. B. Modeling drug albumin binding affinity with E-State topological structure representation. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (6), 2120-2128.
178. Huuskonen, J. QSAR modeling with the electrotopological state indices: predicting the toxicity of organic chemicals. *Chemosphere* **2003,** *50* (7), 949-953.
179. Rose, K.; Hall, L. H. E-state modeling of fish toxicity independent of 3D structure information. *SAR QSAR Environ. Res.* **2003,** *14* (2), 113-129.
180. Roy, K.; Chakraborty, S.; Saha, A. Exploring selectivity requirements for COX-2 versus COX-1 binding of 3,4-diaryloxazolones using E-state index. *Biorg. Med. Chem. Lett.* **2003,** *13* (21), 3753-3757.
181. Chakraborty, S.; Sengupta, C.; Roy, K. Exploring QSAR with E-state index: selectivity requirements for COX-2 versus COX-1 binding of terphenyl methyl sulfones and sulfonamides. *Biorg. Med. Chem. Lett.* **2004,** *14* (18), 4665-4670.
182. Hu, Q. N.; Liang, Y. Z.; Yin, H.; Peng, X. L.; Fang, K. T. Structural interpretation of the topological index. 2. The molecular connectivity index, the Kappa index, and the atom-type E-State index. *J. Chem. Inf. Comp. Sci.* **2004,** *44* (4), 1193-1201.
183. Sengupta, C.; Leonard, J. T.; Roy, K. Exploring QSAR of melatonin receptor ligand benzofuran derivatives using E-state index. *Biorg. Med. Chem. Lett.* **2004,** *14* (13), 3435-3439.
184. Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Hall, L. M. Prediction of aqueous solubility based on large datasets using several QSPR models utilizing topological structure representation. *Chem. Biodiversity* **2004,** *1* (11), 1829-1841.
185. Cash, G. G.; Anderson, B.; Mayo, K.; Bogaczyk, S.; Tunkel, J. Predicting genotoxicity of aromatic and heteroaromatic amines using electrotopological state indices. *Mut. Res. Gen. Toxicol. Environ. Mutagen.* **2005,** *585* (1-2), 170-183.
186. Mukherjee, S.; Mukherjee, A.; Saha, A. QSAR studies with E-State index: Predicting pharmacophore signals for estrogen receptor binding affinity of triphenylacrylonitriles. *Biol. Pharmaceut. Bull.* **2005,** *28* (1), 154-157.
187. Mukherjee, S.; Mukherjee, A.; Saha, A. QSAR modeling on binding affinity of diverse estrogenic flavonoids: electronic, topological and spatial functions in quantitative approximation. *Theochem. J. Mol. Struct.* **2005,** *715* (1-3), 85-90.
188. Roy, K.; Ghosh, G. QSTR with extended topochemical atom indices. Part 5: Modeling of the acute toxicity of phenylsulfonyl carboxylates to Vibrio fischeri using genetic function approximation. *Biorg. Med. Chem.* **2005,** *13* (4), 1185-1194.
189. Liu, S. S.; Yan, D. Q.; Cui, S. H.; Wang, L. S. VSMP for modeling the biodegradability of substituted benzenes based on electrotopological state indices for atom types. *Chin. J. Chem.* **2005,** *23* (5), 622-626.
190. Liu, S. S.; Cui, S. H.; Yin, D. Q.; Shi, Y. Y.; Wang, L. S. QSAR studies on the COX-2 inhibition by 3,4-diarylcycloxazolones based on MEDV descriptor. *Chin. J. Chem.* **2003,** *21* (11), 1510-1516.
191. Liu, S. S.; Yin, C. S.; Wang, L. S. MEDV-13 for QSRR of 62 polychlorinated naphthalenes. *Chin. Chem. Lett.* **2002,** *13* (8), 791-794.
192. Liu, S. S.; Yin, C. S.; Wang, L. S. Combined MEDV-GA-MLR method for QSAR of three panels of steroids, dipeptides, and COX-2 inhibitors. *J. Chem. Inf. Comp. Sci.* **2002,** *42* (3), 749-756.
193. Liu, S. S.; Yin, C. S.; Shi, Y. Y.; Cai, S. X.; Li, Z. L. MEDV-13 for QSAR studies on the COX-2 inhibition by indomethacin amides and esters. *Chin. J. Chem.* **2001,** *19* (8), 751-756.

194. Liu, S. S.; Yin, C. S.; Li, Z. L.; Cai, S. X. QSAR study of steroid benchmark and dipeptides based on MEDV-13. *J. Chem. Inf. Comp. Sci.* **2001,** *41* (2), 321-329.
195. Kellogg, G. E.; Abraham, D. J. Development of empirical molecular interaction models that incorporate hydrophobicity and hydropathy. The HINT paradigm. *Analysis* **1999,** *27* (1), 19-23.
196. Kellogg, G. E. Finding optimum field models for 3D QSAR. *Med. Chem. Res.* **1997,** *7* (6-7), 417-427.
197. Fontaine, F.; Pastor, M.; Zamora, I.; Sanz, F. Anchor-GRIND: Filling the gap between standard 3D QSAR and the GRid-INdependent Descriptors. *J. Med. Chem.* **2005,** *48* (7), 2687-2694.
198. Fontaine, F.; Pastor, M.; Sanz, F. Incorporating molecular shape into the alignment-free GRid-INdependent Descriptors. *J. Med. Chem.* **2004,** *47* (11), 2805-2815.
199. Sciabola, S.; Carosati, E.; Baroni, M.; Mannhold, R. Comparison of ligand-based and structure-based 3D-QSAR approaches: A case study on (aryl-)bridged 2-aminobenzonitriles inhibiting HIV-1 reverse transcriptase. *J. Med. Chem.* **2005,** *48* (11), 3756-3767.
200. Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-organizing molecular field analysis: A tool for structure-activity studies. *J. Med. Chem.* **1999,** *42* (4), 573-583.
201. Kotani, T.; Higashiura, K. Comparative molecular active site analysis (CoMASA). 1. An approach to rapid evaluation of 3D QSAR. *J. Med. Chem.* **2004,** *47* (11), 2732-2742.
202. Polanski, J.; Walczak, B. The comparative molecular surface analysis (COMSA): a novel tool for molecular design. *Comput. Chem.* **2000,** *24* (5), 615-625.
203. Polanski, J.; Gieleciak, R.; Bak, A. The comparative molecular surface analysis (COMSA) - A nongrid 3D QSAR method by a coupled neural network and PLS system: Predicting pK(a) values of benzoic and alkanoic acids. *J. Chem. Inf. Comp. Sci.* **2002,** *42* (2), 184-191.
204. Polanski, J.; Gieleciak, R. The comparative molecular surface analysis (CoMSA) with modified uniformative variable elimination-PLS (UVE-PLS) method: Application to the steroids binding the aromatase enzyme. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (2), 656-666.
205. Polanski, J.; Gieleciak, R.; Wyszomirski, M. Comparative molecular surface analysis (CoMSA) for modeling dye-fiber affinities of the azo and anthraquinone dyes. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (6), 1754-1762.
206. Polanski, J.; Gieleciak, R.; Wyszomirski, M. Mapping dye pharmacophores by the comparative molecular surface analysis (CoMSA): application to the heterocyclic monoazo dyes. *Dyes and Pigm.* **2004,** *62* (1), 61-76.
207. Hasegawa, K.; Morikami, K.; Shiratori, Y.; Ohtsuka, T.; Aoki, Y.; Shimma, N. 3D-QSAR study of antifungal N-myristoyltransferase inhibitors by comparative molecular surface analysis. *Chemom. Intell. Lab. Syst.* **2003,** *69* (1-2), 51-59.
208. Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. Multi-way PLS modeling of structure-activity data by incorporating electrostatic and lipophilic potentials on molecular surface. *Comp. Biol. Chem.* **2003,** *27* (3), 381-386.
209. Hasegawa, K.; Matsuoka, S.; Arakawa, M.; Funatsu, K. New molecular surface-based 3D-QSAR method using Kohonen neural network and 3-way PLS. *Comput. Chem.* **2002,** *26* (6), 583-589.

210. So, S. S.; Karplus, M. Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 2. Applications. *J. Med. Chem.* **1997,** *40* (26), 4360-4371.
211. So, S. S.; Karplus, M. Three-dimensional quantitative structure-activity relationships from molecular similarity matrices and genetic neural networks. 1. Method and validations. *J. Med. Chem.* **1997,** *40* (26), 4347-4359.
212. Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules To Correlate and predict Their Biological Activity. *J. Med. Chem.* **1994,** *37* (24), 4130-4146.
213. Good, A. C.; So, S.-S.; Richards, W. G. Structure-Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993,** *36* (4), 433-438.
214. Carbó-Dorca, R.; Leyda, L.; Arnau, M. How Similar is a Molecule to Another? An Electron Density Measure of Similarity Between Two Molecular Structures. *Int. J. Quant. Chem.* **1980,** *17*, 1185-1189.
215. Good, A. C.; Hodgkin, E. E.; Richards, W. G. Utilization of Gaussian Function for the Rapid Evaluation of Molecular Similarity. *J. Chem. Inf. Comp. Sci.* **1992,** *32* (3), 188-191.
216. Allen, M. S.; LaLoggia, A. J.; Dorn, L. J.; Martin, M. J.; Costantino, G.; Hagen, T. J.; Koehler, K. F.; Skolnick, P.; Cook, J. M. Predictive Binding of Beta-Carboline Inverse Agonists and Antagonists Via the Comfa Golpe Approach. *J. Med. Chem.* **1992,** *35* (22), 4001-4010.
217. Benigni, R.; Cottaramusino, M.; Giorgi, F.; Gallo, G. Molecular Similarity-Matrices and Quantitative Structure-Activity-Relationships - A Case-Study with Methodological Implications. *J. Med. Chem.* **1995,** *38* (4), 629-635.
218. Good, A. C.; Peterson, S. J.; Richards, W. G. Qsars from Similarity-Matrices - Technique Validation and Application in the Comparison of Different Similarity Evaluation Methods. *J. Med. Chem.* **1993,** *36* (20), 2929-2937.
219. Good, A. C.; So, S.-S.; Richards, W. G. Structure-Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993,** *36* (4), 433-438.
220. Robert, D.; Amat, L.; Carbo-Dorca, R. Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures: Prediction of the corticosteroid-binding globulin binding affinity for a steroid family. *J. Chem. Inf. Comp. Sci.* **1999,** *39* (2), 333-344.
221. Fradera, X.; Amat, L.; Besalu, E.; CarboDorca, R. Application of molecular quantum similarity to QSAR. *Quant. Struct. -Act. Relat.* **1997,** *16* (1), 25-32.
222. Bultinck, P.; Carbo-Dorca, R. Molecular quantum similarity matrix based clustering of molecules using dendrograms. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (1), 170-177.
223. Girones, X.; Carbo-Dorca, R. Molecular Quantum Similarity-based QSARs for binding affinities of several steroid sets. *J. Chem. Inf. Comp. Sci.* **2002,** *42* (5), 1185-1193.
224. Girones, X.; Carbo-Dorca, R. Using molecular quantum similarity measures under stochastic transformation to describe physical properties of molecular systems. *J. Chem. Inf. Comp. Sci.* **2002,** *42* (2), 317-325.
225. Girones, X.; Gallegos, A.; Carbo-Dorca, R. Antimalarial activity of synthetic 1,2,4-trioxanes and cyclic peroxy ketals, a quantum similarity study. *J. Comput. -Aided Mol. Des.* **2001,** *15* (12), 1053-1063.

226. Girones, X.; Amat, L.; Carbo-Dorca, R. Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships. *SAR QSAR Environ. Res.* **1999,** *10* (6), 545-556.
227. Amat, L.; Robert, D.; Besalu, E.; Carbo-Dorca, R. Molecular quantum similarity measures tuned 3D QSAR: An antitumoral family validation study. *J. Chem. Inf. Comp. Sci.* **1998,** *38* (4), 624-631.
228. Amat, L.; Carbo-Dorca, R.; Ponec, R. Molecular quantum similarity measures as an alternative to log p values in QSAR studies. *J. Comp. Chem.* **1998,** *19* (14), 1575-1583.
229. Girones, X.; Carbo-Dorca, R.; Ponec, R. Molecular basis of LFER. Modeling of the electronic substituent effect using fragment quantum self-similarity measures. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (6), 2033-2038.
230. Amat, L.; Carbo-Dorca, R.; Ponec, R. Simple linear QSAR models based on quantum similarity measures. *J. Med. Chem.* **1999,** *42* (25), 5169-5180.
231. Ponec, R.; Amat, L.; Carbo-Dorca, R. Molecular basis of quantitative structure-properties relationships (QSPR): A quantum similarity approach. *J. Comput. -Aided Mol. Des.* **1999,** *13* (3), 259-270.
232. Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996,** *96* (3), 1027-1044.
233. Sullivan, J. J.; Jones, A. D.; Tanji, K. K. QSAR treatment of electronic substituent effects using frontier orbital theory and topological parameters. *J. Chem. Inf. Comp. Sci.* **2000,** *40* (5), 1113-1127.
234. Zhang, H. Y.; Sun, Y. M.; Chen, D. Z. O-H bond dissociation energies of phenolic compounds are determined by field/inductive effect or resonance effect? A DFT study and its implication. *Quant. Struct. -Act. Relat.* **2001,** *20* (2), 148-152.
235. Delaere, D.; Nguyen, M. T.; Vanquickenborne, L. G. Structure-property relationships in phosphole-containing pi-conjugated systems: A quantum chemical study. *J. Phys. Chem. A* **2003,** *107* (6), 838-846.
236. Singh, P. P.; Pasha, F. A.; Srivastava, H. K. DFT based atomic softness and its application in site selectivity. *QSAR Comb. Sci.* **2003,** *22* (8), 843-851.
237. Singh, P. P.; Srivastava, H. K.; Pasha, F. A. DFT-based QSAR study of testosterone and its derivatives. *Biorg. Med. Chem.* **2004,** *12* (1), 171-177.
238. Wan, J.; Zhang, L.; Yang, G. F. Quantitative structure-activity relationships for phenyl triazolinones of protoporphyrinogen oxidase inhibitors: A density functional theory study. *J. Comp. Chem.* **2004,** *25* (15), 1827-1832.
239. Wan, J.; Zhang, L.; Yang, G. F.; Zhan, C. G. Quantitative structure-activity relationship for cyclic imide derivatives of protoporphyrinogen oxidase inhibitors: A study of quantum chemical descriptors from density functional theory. *J. Chem. Inf. Comp. Sci.* **2004,** *44* (6), 2099-2105.
240. Zhai, Z. C.; Wang, Z. Y.; Wang, L. S. Quantitative structure-property relationship study of GC retention indices for PCDFs by DFT and relative position of chlorine substitution. *Theochem. J. Mol. Struct.* **2005,** *724* (1-3), 115-124.
241. Kikuchi, O. Systematic Qsar Procedures with Quantum Chemical Descriptors. *Quant. Struct. -Act. Relat.* **1987,** *6* (4), 179-184.
242. Karabunarliev, S.; Mekenyan, O. G.; Karcher, W.; Russom, C. L.; Bradbury, S. P. Quantum-chemical descriptors for estimating the acute toxicity of substituted benzenes to the guppy (Poecilia reticulata) and fathead minnow (Pimephales promelas). *Quant. Struct. -Act. Relat.* **1996,** *15* (4), 311-320.

243. Karabunarliev, S.; Mekenyan, O. G.; Karcher, W.; Russom, C. L.; Bradbury, S. P. Quantum-chemical descriptors for estimating the acute toxicity of electrophiles to the fathead minnow (Pimephales promelas): An analysis based on molecular mechanisms. *Quant. Struct. -Act. Relat.* **1996,** *15* (4), 302-310.
244. Chen, J. W.; Wang, L. S. Using MTLSER model and AM1 Hamiltonian in Quantitative Structure-Activity Relationship studies of alkyl(1-phenylsulfonyl)cycloalkane-carboxylates. *Chemosphere* **1997,** *35* (3), 623-631.
245. Chen, J. W.; Peijnenburg, W. J. G. M.; Wang, L. S. Using PM3 Hamiltonian, factor analysis and regression analysis in developing quantitative structure-property relationships for photohydrolysis quantum yields of substituted aromatic halides. *Chemosphere* **1998,** *36* (13), 2833-2853.
246. Chen, J. W.; Peijnenburg, W. J. G. M.; Quan, X.; Zhao, Y. Z.; Xue, D. M.; Yang, F. L. The application of quantum chemical and statistical technique in developing quantitative structure-property relationships for the photohydrolysis quantum yields of substituted aromatic halides. *Chemosphere* **1998,** *37* (6), 1169-1186.
247. Maran, U.; Karelson, M.; Katritzky, A. R. A comprehensive QSAR treatment of the genotoxicity of heteroaromatic and aromatic amines. *Quant. Struct. -Act. Relat.* **1999,** *18* (1), 3-10.
248. Trohalaki, S.; Gifford, E.; Pachter, R. Improved QSARs for predictive toxicology of halogenated hydrocarbons. *Comput. Chem.* **2000,** *24* (3-4), 421-427.
249. Cronin, M. T. D.; Manga, N.; Seward, J. R.; Sinks, G. D.; Schultz, T. W. Parametrization of electrophilicity for the prediction of the toxicity of aromatic compounds. *Chem. Res. Toxicol.* **2001,** *14* (11), 1498-1505.
250. Yourtee, D.; Holder, A. J.; Smith, R.; Morrill, J. A.; Kostoryz, E.; Brockmann, W.; Glaros, A.; Chappelow, C.; Eick, D. Quantum mechanical quantitative structure activity relationships to avoid mutagenicity in dental monomers. *J. Biomater. Sci. , Polym. Ed.* **2001,** *12* (1), 89-105.
251. Borges, E. G.; Takahata, Y. The 4-indolyl-2-guanidinothiazoles QSAR study of antiulcer activity using quantum descriptors. *Theochem. J. Mol. Struct.* **2002,** *580*, 263-270.
252. Tho, I.; Anderssen, E.; Dyrstad, K.; Kleinebudde, P.; Sande, S. A. Quantum chemical descriptors in the formulation of pectin pellets produced by extrusion/spheronisation. *Eur. J. Pharm. Sci.* **2002,** *16* (3), 143-149.
253. Yamagami, C.; Motohashi, N.; Akamatsu, M. Quantum chemical- and 3-D-QSAR (CoMFA) studies of benzalacetones and 1,1,1-trifluoro-4-phenyl-3-buten-2-ones. *Biorg. Med. Chem. Lett.* **2002,** *12* (17), 2281-2285.
254. Chen, J. W.; Xue, X. Y.; Schramm, K. W.; Quan, M.; Yang, F. L.; Kettrup, A. Quantitative structure-property relationships for octanol-air partition coefficients of polychlorinated naphthalenes, chlorobenzenes and p,p '-DDT. *Comp. Biol. Chem.* **2003,** *27* (3), 165-171.
255. Chen, J. W.; Yang, P.; Chen, S.; Quan, X.; Yuan, X.; Schraam, K. W.; Kettrup, A. Quantitative structure-property relationships for vapor pressures of polybrominated diphenyl ethers. *SAR QSAR Environ. Res.* **2003,** *14* (2), 97-111.
256. de Leval, X.; Ilies, M.; Casini, A.; Dogne, J. M.; Scozzafava, A.; Masini, E.; Mincione, F.; Starnotti, M.; Supuran, C. T. Carbonic anhydrase inhibitors: Synthesis and topical intraocular pressure lowering effects of fluorine-containing inhibitors devoid of enhanced reactivity. *J. Med. Chem.* **2004,** *47* (11), 2796-2804.

257. Pankratov, A. N.; Shalabai, A. V. Quantum-chemical evaluation of the protolytic properties of thiophenols. *J. Struct. Chem.* **2004,** *45* (5), 756-761.
258. Safarpour, M. A.; Hemmateenejad, B.; Miri, R.; Jamali, M. Quantum chemical-QSAR study of some newly synthesized 1,4-dihydropyridine calcium channel blockers. *QSAR Comb. Sci.* **2004,** *22* (9-10), 997-1005.
259. Soriano, E.; Cerdan, S.; Ballesteros, P. Computational determination of pK(a) values. A comparison of different theoretical approaches and a novel procedure. *Theochem. J. Mol. Struct.* **2004,** *684* (1-3), 121-128.
260. Staikova, M.; Wania, F.; Donaldson, D. J. Molecular polarizability as a single-parameter predictor of vapour pressures and octanol-air partitioning coefficients of non-polar compounds: a priori approach and results. *Atmos. Environ.* **2004,** *38* (2), 213-225.
261. Zhokhova, N. I.; Baskin, I. I.; Palyulin, V. A.; Zefirov, A. N.; Zefirov, N. S. Fragmental descriptors in QSPR: application to molecular polarizability calculations. *Russ. Chem. Bull.* **2003,** *52* (5), 1061-1065.
262. Verma, R. P.; Kurup, A.; Hansch, C. On the role of polarizability in QSAR. *Biorg. Med. Chem.* **2005,** *13* (1), 237-255.
263. Verma, R. P.; Hansch, C. A comparison between two polarizability parameters in chemical-biological interactions. *Biorg. Med. Chem.* **2005,** *13* (7), 2355-2372.
264. Broto, P.; Moreau, G.; Vandycke, C. Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies. Autocorrelation Descriptor. *Eur. J. Med. Chem.* **1984,** *19*, 66-70.
265. Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995,** *117* (29), 7769-7775.
266. Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *J. Comput. -Aided Mol. Des.* **1997,** *11* (1), 79-92.
267. Silverman, B. D.; Platt, D. E. Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition. *J. Med. Chem.* **1996,** *39* (11), 2129-2140.
268. Silverman, B. D. The Thirty-one Benchmark Steroid Revisited: Comparative Molecular Moment Analysis (CoMMA) with Principal Component Regression. *Quant. Struct. -Act. Relat.* **2000,** *19* (3), 237-246.
269. Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of novel infrared range vibration-based decriptor (EVA) for QSAR studies: 1. General Application. *J. Comput. -Aided Mol. Des.* **1997,** *11* (4), 409-422.
270. Turner, D. B.; Willett, P. The EVA spectral descriptor. *Eur. J. Med. Chem.* **2000,** *35* (4), 367-375.
271. Tuppurainen, K. EEVA (Electronic Eigenvalue): A New QSAR/QSPR Descriptor Based on Molecular Orbital Energies. *SAR QSAR Environ. Res.* **1999,** *10* (1), 39-46.
272. Tuppurainen, K.; Viisas, M.; Laatikainen, R.; Peräkylä, M. Evaluation of Novel Electronic Eigenvalue (EEVA) Molecular Descriptor for QSAR/QSPR Studies: Validation Using a Benchmar Steroid Data Set. *J. Chem. Inf. Comp. Sci.* **2002,** *42* (3), 607-613.
273. Tuppurainen, K.; Ruuskanen, J. Electronic eigenvalue (EEVA): a new QSAR/QSPR descriptor for electronic substituent effects based on molecular orbital ener-

gies. A QSAR approach to the Ah receptor binding affinity of polychlorinated biphenyls (PCBs), dibenzo-p-dioxins (PCDDs) and dibenzofurans (PCDFs). *Chemosphere* **2000,** *41* (6), 843-848.

274. Afzelius, L.; Masimirembwa, C. M.; Karlen, A.; Andersson, T. B.; Zamora, I. Discriminant and quantitative PLS analysis of competitive CYP2C9 inhibitors versus non-inhibitors using alignment independent GRIND descriptors. *J. Comput. -Aided Mol. Des.* **2002,** *16* (7), 443-458.
275. Benedetti, P.; Mannhold, R.; Cruciani, G.; Pastor, M. GBR compounds and mepyramines as cocaine abuse therapeutics: Chemometric studies on selectivity using grid independent descriptors (GRIND). *J. Med. Chem.* **2002,** *45* (8), 1577-1584.
276. Benedetti, P.; Mannhold, R.; Cruciani, G.; Ottaviani, G. GRIND/ALMOND investigations on CysLT(1) receptor antagonists of the quinolinyl(bridged)aryl type. *Biorg. Med. Chem.* **2004,** *12* (13), 3607-3617.
277. Clementi, M.; Clementi, S.; Cruciani, G.; Pastor, M.; Davis, A. M.; Flower, D. R. Robust multivariate statistics and the prediction of protein secondary structure content. *Protein Eng.* **1997,** *10* (7), 747-749.
278. Jain, A. N.; Diettrich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E. Jr.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. Compass: A Shape-based machine learning tool for drug design. *J. Comput. -Aided Mol. Des.* **1994,** *8*, 635-652.
279. Jain, A. N.; Koile, K.; Chapman, D. Compass: Predicting Biological Activities from Molecular Surface Properties. *J. Med. Chem.* **1994,** *37* (15), 2315-2327.
280. Di Marzio, W.; Galassi, S.; Todeschini, R.; Consolaro, F. Traditional versus WHIM molecular descriptors in QSAR approaches applied to fish toxicity studies. *Chemosphere* **2001,** *44* (3), 401-406.
281. Bravi, G.; Wikel, J. H. Application of MS-WHIM descriptors: 1. Introduction of new molecular surface properties and 2. Prediction of binding affinity data. *Quant. Struct. -Act. Relat.* **2000,** *19* (1), 29-38.
282. Menezes, F. A. S.; Montanari, C. A.; Bruns, R. E. 3D-WHIM pattern recognition study for bisamidines. A structure-property relationship study. *J. Braz. Chem. Soc.* **2000,** *11* (4), 393-397.
283. Gramatica, P.; Navas, N.; Todeschini, R. 3D-modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physicochemical properties of polychlorinated biphenyls (PCBs). *Chemom. Intell. Lab. Syst.* **1998,** *40* (1), 53-63.
284. Todeschini, R.; Moro, G.; Boggia, R.; Bonati, L.; Cosentino, U.; Lasagni, M.; Pitea, D. Modeling and prediction of molecular properties. Theory of grid-weighted holistic invariant molecular (G-WHIM) descriptors. *Chemom. Intell. Lab. Syst.* **1997,** *36* (1), 65-73.
285. Todeschini, R.; Bettiol, C.; Giurin, G.; Gramatica, P.; Miana, P.; Argese, E. Modeling and prediction by using WHIM descriptors in QSAR studies: Submitochondrial particles (SMP) as toxicity biosensors of chlorophenols. *Chemosphere* **1996,** *33* (1), 71-79.
286. Todeschini, R.; Vighi, M.; Provenzani, R.; Finizio, A.; Gramatica, P. Modeling and prediction by using WHIM descriptors in QSAR studies: Toxicity of heterogeneous chemicals on Daphnia magna. *Chemosphere* **1996,** *32* (8), 1527-1545.
287. Silverman, B. D. Three-dimensional moments of molecular property fields. *J. Chem. Inf. Comp. Sci.* **2000,** *40* (6), 1470-1476.

288. Turner, D. B.; Willett, P. Evaluation of the EVA descriptor for QSAR studies: 3. The use of a genetic algorithm to search for models with enhanced predictive properties (EVA_GA). *J. Comput. -Aided Mol. Des.* **2000,** *14* (1), 1-21.
289. Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Evaluation of a novel molecular vibration-based descriptor (EVA) for QSAR studies: 2. Model validation using a benchmark steroid dataset. *J. Comput. -Aided Mol. Des.* **1999,** *13* (3), 271-296.
290. Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B. Q.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997,** *119* (43), 10509-10524.
291. Vedani, A.; Dobler, M. Multidimensional QSAR: Moving from three- to five-dimensional concepts. *Quant. Struct. -Act. Relat.* **2002,** *21* (4), 382-390.
292. Albuquerque, M. G.; Hopfinger, A. J.; Barreiro, E. J.; de Alencastro, R. B. Four-dimensional quantitative structure-activity relationship analysis of a series of interphenylene 7-oxabicycloheptane oxazole thromboxane A(2) receptor antagonists. *J. Chem. Inf. Comp. Sci.* **1998,** *38* (5), 925-938.
293. Klein, C. D. P.; Hopfinger, A. J. Pharmacological activity and membrane interactions of antiarrhythmics: 4D-QSAR/QSPR analysis. *Pharm. Res.* **1998,** *15* (2), 303-311.
294. Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three and four dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2D6 inhibitors. *Pharmacogenetics* **1999,** *9* (4), 477-489.
295. Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three- and four-dimensional quantitative structure activity relationship analyses of cytochrome P-450 3A4 inhibitors. *J. Pharmacol. Exp. Ther.* **1999,** *290* (1), 429-438.
296. Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three- and four-dimensional-quantitative structure activity relationship (3D/4D-QSAR) analyses of CYP2C9 inhibitors. *Drug Metab. Dispos.* **2000,** *28* (8), 994-1002.
297. Duca, J. S.; Tseng, Y. F.; Hopfinger, A. J. 4D-QSPR analysis and virtual screening in materials science. *Adv. Mater.* **2001,** *13* (22), 1713-1717.
298. Ravi, M.; Hopfinger, A. J.; Hormann, R. E.; Dinan, L. 4D-QSAR analysis of a set of ecdysteroids and a comparison to CoMFA modeling. *J. Chem. Inf. Comp. Sci.* **2001,** *41* (6), 1587-1604.
299. Streich, D.; Neuburger-Zehnder, M.; Vedani, A. Induced fit - The key for understanding LSD activity? A 4D-QSAR study on the 5-HT2A receptor system. *Quant. Struct. -Act. Relat.* **2001,** *19* (6), 565-573.
300. Krasowski, M. D.; Hong, X. A.; Hopfinger, A. J.; Harrison, N. L. 4D-QSAR analysis of a set of propofol analogues: Mapping binding sites for an anesthetic phenol on the GABA(A) receptor. *J. Med. Chem.* **2002,** *45* (15), 3210-3221.
301. Kuz'min, V. E.; Artemenko, A. G.; Lozitsky, V. P.; Muratov, E. N.; Fedtchouk, A. S.; Dyachenko, N. S.; Nosach, L. N.; Gridina, T. L.; Shitikova, L. I.; Mudrik, L. M.; Mescheriakov, A. K.; Chelombitko, V. A.; Zheltvay, A. I.; Vanden Eynde, J. J. The analysis of structure-anticancer and antiviral activity relationships for macrocyclic pyridinophanes and their analogues on the basis of 4D QSAR models (simplex representation of molecular structure). *Acta Biochim. Pol.* **2002,** *49* (1), 157-168.

302. Santos, O. A.; Hopfinger, A. J. The 4D-QSAR paradigm: Application to a novel set of nonpeptidic HIV protease inhibitors. *Quant. Struct. -Act. Relat.* **2002,** *21* (4), 369-381.
303. Hong, X.; Hopfinger, A. J. 3D-pharmacophores of flavonoid binding at the benzodiazepine GABA(A) receptor site using 4D-QSAR analysis. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (1), 324-336.
304. Liu, J. Z.; Pan, D. H.; Tseng, Y. F.; Hopfinger, A. J. 4D-QSAR analysis of a series of antifungal P450 inhibitors and 3D-pharmacophore comparisons as a function of alignment. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (6), 2170-2179.
305. Lukacova, V.; Balaz, S. Multimode ligand binding in receptor site modeling: Implementation in CoMFA. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (6), 2093-2105.
306. Polanski, J.; Bak, A. Modeling steric and electronic effects in 3D-and 4D-QSAR schemes: Predicting benzoic pK(a) values and steroid CBG binding affinities. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (6), 2081-2092.
307. Senese, C. L.; Hopfinger, A. J. A simple clustering technique to improve QSAR model selection and predictivity: Application to a receptor independent 4D-QSAR analysis of cyclic urea derived inhibitors of HIV-1 protease. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (6), 2180-2193.
308. Pasqualoto, K. F. M.; Ferreira, E. I.; Santos, O. A.; Hopfinger, A. J. Rational design of new antituberculosis agents: Receptor-independent four-dimensional quantitative structure-activity relationship analysis of a set of isoniazid derivatives. *J. Med. Chem.* **2004,** *47* (15), 3755-3764.
309. Polanski, J.; Bak, A.; Gieleciak, R.; Magdziarz, T. Self-organizing neural networks for Modeling robust 3D and 4D QSAR: Application to dihydrofolate reductase inhibitors. *Molecules* **2004,** *9* (12), 1148-1159.
310. da Cunha, E. F. F.; Albuquerque, M. G.; Antunes, O. A. C.; de Alencastro, R. B. 4D-QSAR models of HOE/BAY-793 analogues as HIV-1 protease inhihitors. *QSAR Comb. Sci.* **2005,** *24* (2), 240-253.
311. Kuz'min, V. E.; Artemenko, A. G.; Lozytska, R. N.; Fedtchouk, A. S.; Lozitsky, V. P.; Muratov, E. N.; Mescheriakov, A. K. Investigation of anticancer activity of macrocyclic Schiff bases by means of 4D-QSAR based on simplex representation of molecular structure. *SAR QSAR Environ. Res.* **2005,** *16* (3), 219-230.
312. Lill, M. A.; Dobler, M.; Vedani, A. In silico prediction of receptor-mediated environmental toxic phenomena - Application to endocrine disruption. *SAR QSAR Environ. Res.* **2005,** *16* (1-2), 149-169.
313. Tsai, K. C.; Lin, T. H. A ligand-based molecular modeling study on some matrix metalloproteinase-1 inhibitors using several 3D QSAR techniques. *J. Chem. Inf. Comp. Sci.* **2004,** *44* (5), 1857-1871.
314. Dobler, M.; Lill, M. A.; Vedani, A. From crystal structures and their analysis to the in silico prediction of toxic phenomena. *Helv. Chim. Acta* **2003,** *86* (5), 1554-1568.
315. Pan, D. H.; Tseng, Y. F.; Hopfinger, A. J. Quantitative structure-based design: Formalism and application of receptor-dependent RD-4D-QSAR analysis to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (5), 1591-1607.
316. Ducki, S.; Mackenzie, G.; Lawrence, N. J.; Snyder, J. P. Quantitative structure-activity relationship (5D-QSAR) study of combretastatin-like analogues as inhibitors of tubulin assembly. *J. Med. Chem.* **2005,** *48* (2), 457-465.

317. Vedani, A.; Dobler, M.; Dollinger, H.; Hasselbach, K. M.; Birke, F.; Lill, M. A. Novel ligands for the chemokine receptor-3 (CCR3): A receptor-modeling study based on 5D-QSAR. *J. Med. Chem.* **2005,** *48* (5), 1515-1527.
318. Vedani, A.; Dobler, M.; Lill, M. A. Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *J. Med. Chem.* **2005,** *48* (11), 3700-3703.
319. Rännar, S.; Lindgren, F.; Geladi, P.; Wold, S. A PLS Kernel Algorithm for data set with many variables and fewer objects. Part 1: Theory and algorithm. *J. Chemom.* **1994,** *8*, 111-125.
320. Geladi, P.; Kowalski, B. R. Partial least-squares regression: A Tutorial. *Anal. Chim. Acta.* **1986,** *185* (1), 1-17.
321. Lindgren, F.; Rännar, S. Alternative partial least squares (PLS). *Perspect. Drug Discov. Design* **1998,** *12/13/14*, 105-113.
322. Ferreira, M. M. C. Multivariate QSAR. *J. Braz. Chem. Soc.* **2002,** *13* (6), 742-753.
323. Takimoto, E.; Koya, S.; Maruoka, A. *Boosting based on divide and merge;* 2004.
324. Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. Detecting "bad" regression models: multicriteria fitness functions in regression analysis. *Anal. Chim. Acta.* **2004,** *515* (1), 199-208.
325. Dong, N.; Lu, W. C.; Chen, N. Y.; Zhu, Y. C.; Chen, K. X. Using support vector classification for SAR of fentanyl derivatives. *Acta Pharm. Sin.* **2005,** *26* (1), 107-112.
326. Figueiredo, L. J. O.; Antunes, O. A. C. Chemometric classification of HIV-1 protease inhibitors. *Int. J. Quant. Chem.* **2000,** *76* (6), 744-755.
327. Gute, B. D.; Basak, S. C. Molecular similarity-based estimation of properties: a comparison of three structure spaces. *J. Mol. Graph. Model.* **2001,** *20* (1), 95-109.
328. Gute, B. D.; Grunwald, G. D.; Mills, D.; Basak, S. C. Molecular similarity based estimation of properties: A comparison of structure spaces and property spaces. *SAR QSAR Environ. Res.* **2001,** *11* (5-6), 363-382.
329. Itskowitz, P.; Tropsha, A. kappa Nearest neighbors QSAR modeling as a variational problem: Theory and applications. *J. Chem. Inf. Mod.* **2005,** *45* (3), 777-785.
330. Jain, B. J.; Wysotzki, F. A k-winner-takes-all classifier for structured data. *Ki 2003: Advances in Artificial Intelligence* **2003,** *2821*, 342-354.
331. Jiang, Y.; Zhou, Z. H. Editing training data for kNN classifiers with neural network ensemble. *Advances in Neural Networks - Isnn 2004, Pt 1* **2004,** *3173*, 356-361.
332. McElroy, N. R.; Jurs, P. C.; Morisseau, C.; Hammock, B. D. QSAR and classification of murine and human soluble epoxide hydrolase inhibition by urea-like compounds. *J. Med. Chem.* **2003,** *46* (6), 1066-1080.
333. Pan, F.; Wang, B. Y.; Hu, X.; Perrizo, W. Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis. *J. Biomed. Inf.* **2004,** *37* (4), 240-248.
334. Petridis, V.; Kaburlasos, V. G. FINkNN: A fuzzy interval number k-nearest neighbor classifier for prediction of sugar production from populations of samples. *J. Mach. Learn. Res.* **2004,** *4* (1), 17-37.
335. Rosa, J. L. A.; Ebecken, N. F. F. Data mining for data classification based on the KNN-fuzzy method supported by genetic algorithm. *High Performance Computing for Computational Science - Vecpar 2002* **2003,** *2565*, 126-133.
336. Shen, M.; Beguin, C.; Golbraikh, A.; Stables, J. P.; Kohn, H.; Tropsha, A. Application of predictive QSAR models to database mining: Identification and experimen-

tal validation of novel anticonvulsant compounds. *J. Med. Chem.* **2004,** *47* (9), 2356-2364.
337. Xiao, Z. Y.; Varma, S.; Xiao, Y. D.; Tropsha, A. Modeling of p38 mitogen-activated protein kinase inhibitors using the Catalyst (TM) HypoGen and k-nearest neighbor QSAR methods. *J. Mol. Graph. Model.* **2004,** *23* (2), 129-138.
338. Zheng, W. F.; Tropsha, A. Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. *J. Chem. Inf. Comp. Sci.* **2000,** *40* (1), 185-194.
339. Mazzatorta, P.; Benfenati, E.; Lorenzini, P.; Vigni, M. QSAR in ecotoxicity: An Overview of modern classification techniques. *J. Chem. Inf. Comp. Sci.* **2004,** *44* (1), 105-112.
340. Burr, T. L.; Fry, H. A. Biased regression: The case for cautious application. *Technometrics* **2005,** *47* (3), 284-296.
341. Farkas, O.; Heberger, K. R. Comparison of ridge regression, partial least-squares, pairwise correlation, forward- and best subset selection methods for prediction of retention indices for aliphatic alcohols. *J. Chem. Inf. Mod.* **2005,** *45* (2), 339-346.
342. Khalaf, G.; Shukur, G. Choosing ridge parameter for regression problems. *Comm. Stat.* **2005,** *34* (5), 1177-1182.
343. Forrester, J. B.; Kalivas, J. H. Ridge regression optimization using a harmonious approach. *J. Chemom.* **2004,** *18* (7-8), 372-384.
344. Jackson, J. E. *A User's Guide to Principal Components;* Wiley & Sons: 1991.
345. Wold, S.; Sjostrom, M. Chemometrics, present and future success. *Chemom. Intell. Lab. Syst.* **1998,** *44* (1-2), 3-14.
346. Wold, S.; Sjostrom, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001,** *58* (2), 109-130.
347. Wold, S.; Josefson, M.; Gottfries, J.; Linusson, A. The utility of multivariate design in PLS modeling. *J. Chemom.* **2004,** *18* (3-4), 156-165.
348. Höskuldsson, A. PLS Regression methods. *J. Chemom.* **1988,** *2*, 211-228.
349. Helland, I. S. On the Structure of Partial Least Squares Regression. *Comm. Stat.* **1988,** *17* (2), 581-607.
350. de Jong, S. SIMPLS: An Alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **1993,** *18*, 251-263.
351. Bush, B. L.; Nachbar, R. B. Jr. Sample-distance Partial Least Squares: PLS optimized for many variables, with application to CoMFA. *J. Comput. -Aided Mol. Des.* **1993,** *7* , 587-619.
352. Wang, T. W.; Khettry, A.; Berry, M.; Batra, J. SVDPLS: An Efficient Algorithm for Performing PLS. 1994.
353. Frank, I. E.; Friedman, J. H. A Statistical View of Some Chemometrics Regression Tools. *Technometrics* **1993,** *35* (2), 109-135.
354. de Jong, S.; Kiers, H. A. L. Principal covariates regression. Part I. Theory. *Chemom. Intell. Lab. Syst.* **1992,** *14* , 155-164.
355. Stone, M.; Brooks, R. J. Continuum Regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinalry Least Squares, Partial Least Squares and Principal Components Regression. *J. R. Statist. Soc. B.* **1990,** *52* (2), 237-269.
356. Wu, W.; Manne, R. Fast regression methods in Lanczos (or PLS-1) basis. Theory and applications. *Chemom. Intell. Lab. Syst.* **2000,** *51* (2), 145-161.

357. Barros, A. S.; Rutledge, D. N. Principal components transform-partial least squares: a novel method to accelerate cross-validation in PLS regression. *Chemom. Intell. Lab. Syst.* **2004,** *73* (2), 245-255.
358. Li, B.; Morris, A. J.; Martin, E. B. Generalized partial least squares regression based on the penalized minimum norm projection. *Chemom. Intell. Lab. Syst.* **2004,** *72* (1), 21-26.
359. Hirst, J. D. Nonlinear quantitative structure-activity relationship for the inhibition of dihydrofolate reductase by pyrimidines. *J. Med. Chem.* **1996,** *39* (18), 3526-3532.
360. Hasegawa, K.; Kimura, T.; Miyashita, Y.; Funatsu, K. Nonlinear partial least squares modeling of phenyl alkylamines with the monoamine oxidase inhibitory activities. *J. Chem. Inf. Comp. Sci.* **1996,** *36* (5), 1025-1029.
361. Hasegawa, K.; Kimura, T.; Funatsu, K. Nonlinear CoMFA using QPLS as a novel 3D-QSAR approach. *Quant. Struct. -Act. Relat.* **1997,** *16* (3), 219-223.
362. Heberger, K.; Borosy, A. P. Comparison of chemometric methods for prediction of rate constants and activation energies of radical addition reactions. *J. Chemom.* **1999,** *13* (3-4), 473-489.
363. Eriksson, L.; Johansson, E.; Lindgren, F.; Wold, S. GIFI-PLS: Modeling of non-linearities and discontinuities in QSAR. *Quant. Struct. -Act. Relat.* **2000,** *19* (4), 345-355.
364. Hasegawa, K.; Funatsu, K. Partial least squares modeling and genetic algorithm optimization in quantitative structure-activity relationships. *SAR QSAR Environ. Res.* **2000,** *11* (3-4), 189-209.
365. Lucic, B.; Nadramija, D.; Basic, I.; Trinajstic, N. Toward generating simpler QSAR models: Nonlinear multivariate regression versus several neural network ensembles and some related methods. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (4), 1094-1102.
366. Yamazaki, K.; Kanaoka, M. Computational prediction of the plasma protein-binding percent of diverse pharmaceutical compounds. *J. Pharm. Sci.* **2004,** *93* (6), 1480-1494.
367. Zhang, H. B. A new approach for the tissue-blood partition coefficients of neutral and ionized compounds. *J. Chem. Inf. Mod.* **2005,** *45* (1), 121-127.
368. Tang, K. L.; Li, T. H. Comparison of different partial least-squares methods in quantitative structure-activity relationships. *Anal. Chim. Acta.* **2003,** *476* (1), 85-92.
369. Wold, S.; Trygg, J.; Berglund, A.; Antti, H. Some recent developments in PLS modeling. *Chemom. Intell. Lab. Syst.* **2001,** *58* (2), 131-150.
370. Constans, P.; Hirst, J. D. Nonparametric regression applied to quantitative structure - Activity relationships. *J. Chem. Inf. Comp. Sci.* **2000,** *40* (2), 452-459.
371. Hirst, J. D.; McNeany, T. J.; Howe, T.; Whitehead, L. Application of non-parametric regression to quantitative structure-activity relationships. *Biorg. Med. Chem.* **2002,** *10* (4), 1037-1041.
372. McNeany, T. J.; Hirst, J. D. Inhibition of the tyrosine kinase, Syk, analyzed by stepwise nonparametric regression. *J. Chem. Inf. Mod.* **2005,** *45* (3), 768-776.
373. Bazoui, H.; Zahouily, M.; Boulajaaj, S.; Sebti, S.; Zakarya, D. QSAR for anti-HIV activity of HEPT derivatives. *SAR QSAR Environ. Res.* **2002,** *13* (6), 567-577.
374. Collantes, E. R.; Gahimer, T.; Welsh, W. J.; Grayson, M. Evaluation of computational chemistry approaches for predicting the properties of polyimides. *Comput. Theor. Polym. Sci.* **1996,** *6* (1-2), 29-40.

375. Czerminski, R.; Yasri, A.; Hartsough, D. Use of Support Vector Machine in pattern classification: Application to QSAR studies. *Quant. Struct. -Act. Relat.* **2001,** *20* (3), 227-240.
376. Funar-Timofei, S.; Suzuki, T.; Paier, J. A.; Steinreiber, A.; Faber, K.; Fabian, W. M. F. Quantitative structure - Activity relationships for the enantioselectivity of oxirane ring-opening catalyzed by epoxide hydrolases. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (3), 934-940.
377. Jalali-Heravi, M.; Fatemi, M. H. Artificial neural network modeling of Kovats retention indices for noncyclic and monocyclic terpenes. *J. Chromatogr. A* **2001,** *915* (1-2), 177-183.
378. Livingstone, D. J.; Manallack, D. T. Neural networks in 3D QSAR. *QSAR Comb. Sci.* **2003,** *22* (5), 510-518.
379. Mazzatorta, P.; Vracko, M.; Benfenati, E. ANVAS: Artificial Neural Variables Adaptation System for descriptor selection. *J. Comput. -Aided Mol. Des.* **2003,** *17* (5-6), 335-346.
380. Salt, D. W.; Yildiz, N.; Livingstone, D. J.; Tinsley, C. J. The Use of Artificial Neural Networks in Qsar. *Pesticide Science* **1992,** *36* (2), 161-170.
381. Song, X. H.; Chen, Z.; Yu, R. Q. Artificial Neural Networks Applied to the Quantitative Structure-Activity Relationship Study of Parasubstituted Phenols. *Sci. Chin. B* **1993,** *36* (12), 1443-1450.
382. Tetko, I. V.; Luik, A. I.; Poda, G. I. Applications of Neural Networks in Structure-Activity-Relationships of A Small Number of Molecules. *J. Med. Chem.* **1993,** *36* (7), 811-814.
383. Tmej, C.; Chiba, P.; Schaper, K. J.; Ecker, G.; Fleischhacker, W. Artificial neural networks as versatile tools for prediction of MDR-modulatory activity. *Adv. Exp. Med. Biol.* **1999,** *457*, 95-105.
384. Turner, J. V.; Maddalena, D. J.; Cutler, D. J. Pharmacokinetic parameter prediction from drug structure using artificial neural networks. *Int. J. Pharm.* **2004,** *270* (1-2), 209-219.
385. Vendrame, R.; Braga, R. S.; Takahata, Y.; Galvao, D. S. Structure-carcinogenic activity relationship studies of polycyclic aromatic hydrocarbons (PAHs) with pattern-recognition methods. *Theochem. J. Mol. Struct.* **2001,** *539*, 253-265.
386. Vracko, M. A study of structure carcinogenic potency relationship with artificial neural networks. The using of descriptors related to geometrical and electronic structures. *J. Chem. Inf. Comp. Sci.* **1997,** *37* (6), 1037-1043.
387. Vracko, M.; Novic, M.; Zupan, J. Study of structure-toxicity relationship by a counterpropagation neural network. *Anal. Chim. Acta.* **1999,** *384* (3), 319-332.
388. Weekes, D.; Fogel, G. B. Evolutionary optimization, backpropagation, and data preparation issues in QSAR modeling of HIV inhibition by HEPT derivatives. *Biosystems* **2003,** *72* (1-2), 149-158.
389. Yan, A. X.; Jiao, G. M.; Hu, Z. D.; Fan, B. T. Use of artificial neural networks to predict the gas chromatographic retention index data of alkylbenzenes on carbowax-20M. *Comput. Chem.* **2000,** *24* (2), 171-179.
390. Zahouily, M.; Rhihil, A.; Bazoui, H.; Sebti, S.; Zakarya, D. Structure-toxicity relationships study of a series of organophosphorus insecticides. *J. Mol. Model.* **2002,** *8* (5), 168-172.
391. Zhang, R. S.; Liu, S. H.; Liu, M. C.; Hu, Z. Neural network molecular descriptors approach to the prediction of properties of alkenes. *Comput. Chem.* **1997,** *21* (5), 335-341.

392. Zupan, J.; Novic, M. Optimisation of structure representation for QSAR studies. *Anal. Chim. Acta.* **1999,** *388* (3), 243-250.
393. Tetko, I. V.; Livingstone, D. J.; Luik, A. I. Neural-Network Studies 1. Comparison of Overfitting and Overtraining. *J. Chem. Inf. Comp. Sci.* **1995,** *35* (5), 826-833.
394. Tominaga, Y. Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN. *Chemom. Intell. Lab. Syst.* **1999,** *49* (1), 105-115.
395. Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S. On the use of neural network ensembles in QSAR and QSPR. *J. Chem. Inf. Comp. Sci.* **2002,** *42* (4), 903-911.
396. Livingstone, D. J.; Salt, D. W. Regression-Analysis for Qsar Using Neural Networks. *Biorg. Med. Chem. Lett.* **1992,** *2* (3), 213-218.
397. Manallack, D. T.; Livingstone, D. J. Relating Biological-Activity to Chemical-Structure Using Neural Networks. *Pesticide Science* **1995,** *45* (2), 167-170.
398. Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design;* 2nd ed.; Wiley-VCH: 1999.
399. Guha, R.; Jurs, P. C. Interpreting computational neural network QSAR models: A measure of descriptor importance. *J. Chem. Inf. Mod.* **2005,** *45* (3), 800-806.
400. Guha, R.; Stanton, D. T.; Jurs, P. C. Interpreting computational neural network quantitative structure-activity relationship models: A detailed interpretation of the weights and biases. *J. Chem. Inf. Mod.* **2005,** *45* (4), 1109-1121.
401. Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Model.* **2002,** *20* (4), 269-276.
402. Bottou, L.; Vapnik, V. Local Learning Algorithms. *Neural Computation* **1992,** *4* (6), 888-900.
403. Bao, Y. G.; Ishii, N.; Du, X. Y. Combining multiple k-nearest neighbor classifiers using different distance functions. *Intelligent Data Engineering and Automated Learning Ideal 2004, Proceedings* **2004,** *3177*, 634-641.
404. Guo, G.; Wang, H.; Bell, D. Similarity-based data reduction techniques. *J. Res. Pract. Inf. Tech.* **2005,** *37* (2), 211-232.
405. Pechenizkiy, M. The impact of feature extraction on the performance of a classifier: kNN, Naive Bayes and C4.5. *Advances in Artificial Intelligence, Proceedings* **2005,** *3501*, 268-279.
406. Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comp. Sci.* **2004,** *44*, 1912-1928.
407. Livingstone, D. J.; Manallack, D. T.; Tetko, I. V. Data modelling with neural networks: Advantages and limitations. *J. Comput. -Aided Mol. Des.* **1997,** *11* (2), 135-142.
408. Ortiz, A. R.; Pastor, M.; Palomer, A.; Cruciani, G.; Gago, F.; Wade, R. C. Reliability of comparative molecular field analysis models: Effects of data scaling and variable selection using a set of human synovial fluid phospholipase A(2) inhibitors. *J. Med. Chem.* **1997,** *40* (7), 1136-1148.
409. Topliss, J. G.; Costello, R. J. Chance correlations in Structure-Activity Studies Using Multiple Regression Analysis. *J. Med. Chem.* **1972,** *15* (10), 1066-1068.
410. Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance on being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *Quant. Struct. -Act. Relat.* **2003,** *22*, 69-76.
411. Thomsen, M.; Dobel, S.; Lassen, P.; Carlsen, L.; Mogensen, B. B.; Hansen, P. E. Reverse quantitative structure-activity relationship for modelling the sorption of

esfenvalerate to dissolved organic matter - A multivariate approach. *Chemosphere* **2002,** *49* (10), 1317-1325.
412. Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Chen, X. Q.; Doweyko, A.; Li, Y. *In silico* ADME/Tox: why models fail. *J. Comput. -Aided Mol. Des.* **2003,** *17* (2 - 4), 83-92.
413. Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear Pls Estimations (Golpe) - An Advanced Chemometric Tool for Handling 3D-Qsar Problems. *Quant. Struct. -Act. Relat.* **1993,** *12* (1), 9-20.
414. Greco, G.; Novellino, E.; Pellecchia, M.; Silipo, C.; Vittoria, A. Effects of Variable Selection on Comfa Coefficient Contour Maps in A Set of Triazines Inhibiting Dhfr. *J. Comput. -Aided Mol. Des.* **1994,** *8* (2), 97-112.
415. Hasegawa, K.; Kimura, T.; Funatsu, K. CA strategy for variable selection in QSAR studies: Enhancement of comparative molecular binding energy analysis by GA-based PLS method. *Quant. Struct. -Act. Relat.* **1999,** *18* (3), 262-272.
416. Kimura, T.; Hasegawa, K.; Funatsu, K. GA strategy for variable selection in QSAR studies: GA-based region selection for CoMFA modeling. *J. Chem. Inf. Comp. Sci.* **1998,** *38* (2), 276-282.
417. Norinder, U. Single and domain mode variable selection in 3D QSAR applications. *J. Chemom.* **1996,** *10* (2), 95-105.
418. Yasri, A.; Hartsough, D. Toward an optimal procedure for variable selection and QSAR model building. *J. Chem. Inf. Comp. Sci.* **2001,** *41* (5), 1218-1227.
419. Ferguson, A. M.; Heritage, T.; Jonathon, P.; Pack, S. E.; Phillips, L.; Rogan, J.; Snaith, P. J. EVA: A new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *J. Comput. -Aided Mol. Des.* **1997,** *11* (2), 143-152.
420. Ford, M.; Phillips, L.; Stevens, A. Optimising the EVA descriptor for prediction of biological activity. *Org. Biomol. Chem.* **2004,** *2* (22), 3301-3311.
421. Barreca, M. L.; Rao, A.; De Luca, L.; Zappala, M.; Gurnari, C.; Monforte, P.; De Clercq, E.; Van Maele, B.; Debyser, Z.; Witvrouw, M.; Briggs, J. M.; Chimirri, A. Efficient 3D database screening for novel HIV-1IN inhibitors. *J. Chem. Inf. Comp. Sci.* **2004,** *44* (4), 1450-1455.
422. Bruno-Blanch, L.; Galvez, J.; Garcia-Domenech, R. Topological virtual screening: A way to find new anticonvulsant drugs from chemical diversity. *Biorg. Med. Chem. Lett.* **2003,** *13* (16), 2749-2754.
423. Clark, D. E.; Higgs, C.; Wren, S. P.; Dyke, H. J.; Wong, M.; Norman, D.; Lockey, P. M.; Roach, A. G. A virtual screening approach to finding novel and potent antagonists at the melanin-concentrating hormone 1 receptor. *J. Med. Chem.* **2004,** *47* (16), 3962-3971.
424. Duart, M. J.; nton-Fos, G. M.; Aleman, P. A.; Gonzalez-Rosende, M. E.; Galvez, J.; Garcia-Domenech, R. New potential antihistaminic compounds. Virtual combinatorial chemistry, computational screening, real synthesis, and pharmacological evaluation. *J. Med. Chem.* **2005,** *48* (4), 1260-1264.
425. Godden, J. W.; Stahura, F. L.; Bajorath, J. POT-DMC: A virtual screening method for the identification of potent hits. *J. Med. Chem.* **2004,** *47* (23), 5608-5611.
426. Alberts, I. L.; Todorov, N. P.; Kallblad, P.; Dean, P. M. Ligand docking and design in a flexible receptor site. *QSAR Comb. Sci.* **2005,** *24* (4), 503-507.
427. Bindewald, E.; Skolnick, J. A scoring function for docking ligands to low-resolution protein structures. *J. Comp. Chem.* **2005,** *26* (4), 374-383.

428. Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L. Comparative study of several algorithms for flexible ligand docking. *J. Comput. -Aided Mol. Des.* **2003,** *17* (11), 755-763.
429. Cavasotto, C. N.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004,** *337* (1), 209-225.
430. Cho, A. E.; Guallar, V.; Berne, B. J.; Friesner, R. Importance of accurate charges in molecular docking: Quantum mechanical/molecular mechanical (QM/MM) approach. *J. Comp. Chem.* **2005,** *26* (9), 915-931.
431. de Magalhaes, C. S.; Barbosa, H. J. C.; Dardenne, L. E. A genetic algorithm for the ligand-protein docking problem. *Genetics and Molecular Biology* **2004,** *27* (4), 605-610.
432. Fahmy, A.; Wagner, G. TreeDock: A tool for protein docking based on minimizing van der Waals energies. *J. Am. Chem. Soc.* **2002,** *124* (7), 1241-1250.
433. Fradera, X.; Knegtel, R. M. A.; Mestres, J. Similarity-driven flexible ligand docking. *Proteins: Struct. Funct. Genet.* **2000,** *40* (4), 623-636.
434. Chipot, C.; Rozanska, X.; Dixit, S. B. Can free energy calculations be fast and accurate at the same time? Binding of low-affinity, non-peptide inhibitors to the SH2 domain of the src protein. *J. Comput. -Aided Mol. Des.* **2005,** *19* (11), 765-770.
435. Ortiz, A. R.; Gomez-Puertas, P.; Leo-Macias, A.; Lopez-Romero, P.; Lopez-Vinas, E.; Morreale, A.; Murcia, M.; Wang, K. Computational approaches to model ligand selectivity in drug design. *Curr. Top. Med. Chem.* **2006,** *6* (1), 41-55.
436. Lazaridis, T. Binding affinity and specificity from computational studies. *Curr. Org. Chem.* **2002,** *6* (14), 1319-1332.
437. Otlewski, J.; Apostoluk, W. Structural and energetic aspects of protein-protein recognition. *Acta Biochim. Pol.* **1997,** *44* (3), 367-387.
438. Sippl, W. Development of biologically active compounds by combining 3D QSAR and structure-based design methods. *J. Comput. -Aided Mol. Des.* **2002,** *16* (11), 825-830.
439. Sippl, W. Binding affinity prediction of novel estrogen receptor ligands using receptor-based 3-D QSAR methods. *Biorg. Med. Chem.* **2002,** *10* (12), 3741-3755.
440. Vedani, A.; McMasters, D. R.; Dobler, M. Multi-conformational ligand representation in 4D-QSAR: Reducing the bias associated with ligand alignment. *Quant. Struct. -Act. Relat.* **2000,** *19* (2), 149-161.
441. Brown, N.; Mckay, B.; Gasteiger, J. Fingal a novel approach to geometric fingerprinting and a comparative study of its application to 3D-QSAR modelling. *QSAR Comb. Sci.* **2005,** *24* (4), 480-484.
442. Hirons, L.; Holliday, J. D.; Jelfs, S. P.; Willett, P.; Gedeck, P. Use of the R-group descriptor for alignment-free QSAR. *QSAR Comb. Sci.* **2005,** *24* (5), 611-619.
443. Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Feng, J.; Zheng, W.; Tropsha, A. QSAR modeling of datasets with enantioselective compounds using chirality sensitive molecular descriptors. *SAR QSAR Environ. Res.* **2005,** *16* (1-2), 93-102.
444. Seel, M.; Turner, D. B.; Willett, P. Effect of parameter variations on the effectiveness of HQSAR analyses. *Quant. Struct. -Act. Relat.* **1999,** *18* (3), 245-252.
445. Kubinyi, H.; Sadowski, J. Qsar, 3D Qsar and Beyond. *Abstr. Paper. Am. Chem. Soc.* **1999,** *217*, U667.
446. Hopfinger, A. J.; Reaka, A.; Venkatarangan, P.; Duca, J. S.; Wang, S. Construction of a virtual nigh throughput screen by 4D-QSAR analysis: Application to a

combinatorial library of glucose inhibitors of glycogen phosphorylase b. *J. Chem. Inf. Comp. Sci.* **1999,** *39* (6), 1151-1160.
447. Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parametrization, and Performance of MMFF94. *J. Comp. Chem.* **1996,** *17* (5-6), 490-519.
448. Halgren, T. A. Merck Molecular Force Field. II. MMFF94 van der Waals and Electrostatic Parameters for Intermolecular Interactions. *J. Comp. Chem.* **1996,** *17* (5-6), 520-552.
449. Halgren, T. A. Merck Molecular Force Field. III. Molecular Geometries and Vibrational Frequencies for MMFF94. *J. Comp. Chem.* **1996,** *17* (5-6), 553-586.
450. Halgren, T. A. Merck Molecular Force Field. IV. Conformational Energies and Geometries for MMFF94. *J. Comp. Chem.* **1996,** *17* (5-6), 587-615.
451. Halgren, T. A. Merck Molecular Force Field. V. Extension of MMFF94 Using Experimental Data, Additional Computational Data, and Empirical Rules. *J. Comp. Chem.* **1996,** *17* (5-6), 616-641.
452. Basch, H.; Ratner, M. A. Reduced basis set for the gold atom in cluster complexes. *J. Comp. Chem.* **2004,** *25* (7), 899-906.
453. Pakiari, A. H.; Solimannejad, M. An alternative ab initio calculation of orbital energy for a given wave function. *Theochem. J. Mol. Struct.* **2002,** *583*, 99-104.
454. Flamant, I.; Fripiat, J. G.; Delhalle, J. Advantages of the Fourier space RHF band structure approach: Application to polyoxymethylene using a distributed basis set of s-type Gaussian functions. *Int. J. Quant. Chem.* **1998,** *70* (4-5), 1045-1054.
455. Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comp. Chem.* **1996,** *17* (14), 1653-1666.
456. Chong, H. C.; Dong, G. O.; Wanchul, S. Flexible Molecular Superposition: Development of a Combined Similarity Index and Application of the Constrained Optimization Technique. *J. Comp. Chem.* **2001,** *22* (8), 888-900.
457. Good, A. C.; Hodgkin, E. E.; Richards, W. G. Utilization of Gaussian Function for the Rapid Evaluation of Molecular Similarity. *J. Chem. Inf. Comp. Sci.* **1992,** *32* (3), 188-191.
458. Coats, E. A. The CoMFA steroids as a Benchmark Dataset for Development of 3D QSAR Methods. *Perspect. Drug Discov. Design* **1998,** *12-14*, 199-213.
459. Gantchev, T. G.; Ali, H.; van Lier, J. Quantitative Structure-Activity Relationship/Comparative Molecular Field Analysis (QSAR/CoMFA) for Receptor-Binding Properties of Halogenated Estratriol Derivatives. *J. Med. Chem.* **1994,** *37*, 4164-4176.
460. Wiese, T. E.; Polin, L. A.; Palomino, E.; Brooks, S. C. Induction of the Estrogen Specific Mitogenic Response of MCF-7 Cells by Selected Analogues of Estradiol-17B: A 3D QSAR Study. *J. Med. Chem.* **1997,** *40*, 3659-3669.
461. *MATLAB. Program for matrix and technical computing,* version 6.5; MathWorks Inc.: 2006.
462. van de Waterbeemd, H.; Testa, B.; Folkers, G. Computer-Assisted Lead Finding and Optimization: Current Tools for Medicinal Chemistry. In *A General View on Similarity and QSAR Studies.*, Kubinyi, H., Ed.; VHCA: Basel, Switzerland, 1997; pp 7-28.
463. Poso, A.; Tuppurainen, K.; Ruuskanen, J.; Gynther, J. Binding of some dioxins and dibenzofurans to the Ah receptor. A QSAR model based on comparative mo-

lecular field analysis (CoMFA). *Theochem. J. Mol. Struct.* **1993,** *282* (3), 259-264.
464. Waller, C. L.; McKinney, J. D. Comparative Molecular Field Analysis of Polyhalogenated Dibenzo-*p*-dioxins, Dibenzofurans, and Biphenyls. *J. Med. Chem.* **1992,** *35* (20), 3660-3666.
465. Korhonen, S.-P.; Tuppurainen, K.; Laatikainen, R.; Peräkylä, M. FLUFF-BALL, A Template-Based Grid-Independent Superposition and QSAR Technique: Validation Using a Benchmark Steroid Data Set. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (6), 1780-1793.
466. Sonnenschein, C.; Soto, A. M. An updated review of environmental estrogen and androgen mimics and antagonists. *J. Steroid Biochem. Mol. Biol.* **1998,** *65*, 143-150.
467. Gray, L. E.; Kelce, W. R.; Wiese, T.; Tyl, R.; Gaido, K.; Cook, J.; Klinefelter, G.; Desaulniers, D.; Wilson, E.; Zacharewiski, T.; Waller, C.; Foster, P.; Laskey, J.; Reel, J.; Giesy, J.; Laws, S.; McLachlan, J.; Breslin, W.; Cooper, R.; Di Giulio, R.; Johnson, R.; Purdy, R.; Mihaich, E.; Safe, S.; Sonnenschein, C.; Welshons, W.; Miller, R.; McMaster, S.; Colborn, T. Endocrine screening methods workshop report: Detection of estrogenic and androgenic hormonal and antihormonal activity for chemicals that act via receptor or steroidogenic enzyme mechanisms. *Reprod. Toxicol.* **1997,** *11*, 719-750.
468. McLachlan, J. A. Environmental signalling: What embryos and evolution teach us about endocrine disrupting chemicals. *Endocr. Rev.* **2001,** *22*, 319-341.
469. Preziosi, P. Endocrine disrupters as environmental signallers: An introduction. *Pure Appl. Chem.* **1998,** *70*, 1617-1631.
470. Pons, M.; Gagne, D.; Nicolas, J. C.; Mehtali, M. A new cellular-model of response to estrogens-A bioluminescent test to characterize (anti)estrogen molecules. *Biotechniques* **1990,** *9*, 450-459.
471. Soto, A. M.; Sonnenschein, C.; Chung, K. L.; Fernandez, M. F.; Olea, N.; Serrano, F. O. The E-SCREEN assay as a tool to identify estrogens: An update on estrogenic environmental pollutants. *Environ. Health Perspect.* **1995,** *103*, 113-122.
472. Reel, J. R.; Lamb, J. C.; Neal, B. H. Survey and assessment of mammalian estrogen biological assays for hazard characterization. *Fundam. Appl. Toxicol.* **1996,** *34*, 288-305.
473. Shelby, M. D.; Newbold, R. R.; Tully, D. B.; Chae, K.; Davis, V. L. Assessing environmental chemicals for estrogenicity using a combination of in vitro and in vivo assays. *Environ. Health Perspect.* **1996,** *104*, 1296-1300.
474. Bolger, R.; Wiese, T. E.; Ervin, K.; Nestich, S.; Checovich, W. Rapid screening of environmental chemicals for estrogen receptor binding capacity. *Environ. Health Perspect.* **1998,** *106*, 551-557.
475. Fang, H.; Tong, W.; Perkins, R.; Soto, A. M.; Prechtl, N. V.; Sheehan, D. M. Quantitative comparisons of in vitro assays for estrogenic activities. *Environ. Health Perspect.* **2000,** *108*, 723-729.
476. Fang, H.; Tong, W.; Welsh, W. J.; Sheehan, D. M. QSAR models in receptor-mediated effects: the nuclear receptor superfamily. *Theochem. J. Mol. Struct.* **2003,** *622* (1-2), 113-125.
477. Schmieder, P. K.; Ankley, G.; Mekenyan, O.; Walker, J. D. Quantitative structure-activity relationship models for prediction of estrogen receptor binding affin-

ity of structurally diverse chemicals. *Environ. Toxicol. Chem.* **2003,** *22*, 1844-1854.
478. Tong, W.; Perkins, R. QSAR Models for Binding of Estrogenic Compounds to Estrogen Receptor α and β Subtypes. *Endocrinology* **1997,** *138* (9), 4022-4025.
479. Tong, W.; Perkins, R.; Strelitz, R.; Collantes, E. R.; Keenan, S.; Welsh, W. J.; Branham, W. S.; Sheehan, D. M. Quantitative Structure-Activity Relationships (QSARs) for estrogen Binding to the Estrogen Receptor: Predictions across Species. *Environ. Health Perspect.* **1997,** *105* (10), 1116-1124.
480. Waller, C. L.; Oprea, T. I.; Chae, K.; Park, H.-K.; Korach, K. S.; Laws, S. C.; Wiese, T. E.; Kelce, W. R.; Grey, L. E. J. Ligand-based indentification of environmental estrogens. *Chem. Res. Toxicol.* **1996,** *9* (8), 1240-1248.
481. Wolohan, P.; Reichert, D. E. CoMFA and docking study of novel estrogen subtype selective ligands. *J. Comput. -Aided Mol. Des.* **2003,** *17* (5), 313-328.
482. Kubinyi, H. Comparative Molecular Field Analysis (CoMFA). In *Encyclopedia of Computational Chemistry*, Wiley & Sons: 1998; pp 3001-3012.
483. Li, M.; Du, L.; Wu, B.; Xia, L. Self-Organizing Molecular Field Analysis on α_{1a}-Adnoceptor Dihydropyridine Antagonists. *Biorg. Med. Chem.* **2003,** *11* (18), 3945-3951.
484. Smith, P. A.; Sorich, M. J.; McKinnon, R. A.; Miners, J. O. Pharmacophore and Quantitative Structure-Activity Relationship Modeling: Complementary Approaches for the Rationalization and Prediction of UDP-Glucuronolsyltransferase 1A4 Substrate Selectivity. *J. Med. Chem.* **2003,** *46* (9), 1617-1626.
485. Asikainen, A. H.; Ruuskanen, J.; Tuppurainen, K. Spectroscopic QSAR Methods and Self-Organizing Molecular Field Analysis for Relating Molecular Structure and Estrogenic Activity. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (6), 1974-1981.
486. Tuppurainen, K.; Viisas, M.; Peräkylä, M.; Laatikainen, R. Ligand intramolecular motions in ligand-protein interaction: ALPHA, a novel dynamic descriptor and a QSAR study with extented steroid benchmark dataset. *J. Comput. -Aided Mol. Des.* **2004,** *18* (3), 175-187.
487. Wu, B.; Li, M.; Jiang, Z. Z.; Xia, L. Design, synthesis and 3D-QSAR study of N-substituted-3-indolyl-acetamide series as α_1-adrenoceptor antagonists. *Chin. J. Org. Chem.* **2004,** *24* (12), 1587-1594.
488. Martinek, T. A.; Ötvös, F.; Dervarics, M.; Fülöp, F. Ligand-Based Prediction of Active Conformation by 3D-QSAR Flexibility Descriptors and Their Application in 3+3D-QSAR Models. *J. Med. Chem.* **2005,** *48*, 3239-3250.
489. Klocker, J.; Wailzer, B.; Buchbauer, G.; Wolschann, P. Aroma Quality Differentiation of Pyrazine Derivatives Using Self-Organizing Molecular Field Analysis and Artificial Neural Network. *J. Agr. Food. Chem.* **2002,** *50* (17), 4069-4075.
490. Constans, P.; Amat, L.; Carbó-Dorca, R. Toward a global maximization of the molecular similarity function: Superposition of two molecules. *J. Comp. Chem.* **1997,** *18* (6), 826-846.
491. Nissink, J. W. M.; Verdonk, M. L.; Kroon, J.; Mietzner, T.; Klebe, G. Superposition of Molecules: Electron Density Fitting by Application of Fourier Transforms. *J. Comp. Chem.* **1997,** *18* (5), 638-645.
492. Parretti, M. F.; Kroemer, R. T.; Rothman, J. H.; Richards, W. G. Alignment of Molecules by the Monte Carlo Optimization of Molecular Similarity Indices. *J. Comp. Chem.* **1997,** *18* (11), 1344-1353.

493. Chen, H.; Zhou, J.; Xie, G. PARM: A Genetic Evoloved Algorithm To Perdict Bioactivity. *J. Chem. Inf. Comp. Sci.* **1998,** *38* (2), 243-250.
494. Amat, L.; Besalu, E.; Carbó-Dorca, R. Identification of Active Molecular Sites Using Quantum-Self-Similarity Measures. *J. Chem. Inf. Comp. Sci.* **2001,** *41* (4), 978-991.
495. Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S. Molecular Field Topology Analysis Method in QSAR Studies of Organic Compounds. *J. Chem. Inf. Comp. Sci.* **2000,** *40* (3), 659-667.
496. Sadler, B. R.; Cho, S. J.; Ishaq, K. S.; Chae, K.; Korach, K. S. Three-dimensional quantitative structure-activity relationship study of nonsteroidal estrogen receptor ligands using the comparative molecular field analysis cross-validated r(2)-guided region selection approach. *J. Med. Chem.* **1998,** *41* (13), 2261-2267.
497. Bradley, M.; Waller, C. L. Polarizability fields for use in three-dimension quantitative structure-activity relationship (3D-QSAR). *J. Chem. Inf. Comp. Sci.* **2001,** *41* (5), 1301-1307.
498. Lewis, D. F. V. The Calculation of Molar Polarizabilities by the CNDO/2 method: Correlation with the hydrophobic parameter, log P. *J. Comp. Chem.* **1989,** *10* (2), 145-151.
499. Korhonen, S.-P.; Tuppurainen, K.; Laatikainen, R.; Peräkylä, M. Improving the performance of SOMFA by use of standard multivariate methods. *SAR QSAR Environ. Res.* **2005,** *16* (6), 567-579.
500. Cho, S. J.; Tropsha, A. Cross-Validated R2-guided Region Selection for Comparative Molecular Field Analysis: A Simple Method to Achieve Consistent Results. *J. Med. Chem.* **1995,** *38* (7), 1060-1066.
501. Zheng, W. F.; Tropsha, A. A novel nonlinear QSAR method based on K-nearest neighbor principle and variable selection. *Abstr. Paper. Am. Chem. Soc.* **1998,** *215*, U512.
502. Shi, L. M.; Fang, H.; Tong, W.; Wu, J.; Perkins, R.; Blair, R. M.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comp. Sci.* **2001,** *41* (1), 186-195.
503. Bayram, E.; Santago, P. I.; Harris, R.; Xiao, Y.-D.; Clauset, A. J.; Schmitt, J. D. Genetic algorithms and self-organizing maps: a powerful combination for modeling complex QSAR and QSPR problems. *J. Comput. -Aided Mol. Des.* **2004,** *18*, 483-493.
504. Bursi, R.; Groen, M. B. Application of (quantitative) structure-activity relationships to progestagens. *Eur. J. Med. Chem.* **2000,** *35*, 787-796.
505. Cherkasov, A. 'Inductive' Descriptors: 10 Successful Years in QSAR. *Curr. Comput. -Aided Drug Des.* **2005,** *1*, 21-42.
506. Clark, R. D. Boosted leave-many-out cross-validation: the effect of training and test set diversity on PLS statistics. *J. Comput. -Aided Mol. Des.* **2003,** *17* (2-4), 265-275.
507. Feng, J.; Lurati, L.; Ouyang, H.; Robinson, T.; Wang, Y.; Yuan, S.; Young, S. S. Predictive toxicology: Benchmarking molecular descriptors and statistical methods. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (5), 1463-1470.
508. Indahl, U. A twist to partial least squares regression. *J. Chemom.* **2005,** *19* (1), 32-44.
509. Asikainen, A. H.; Tuppurainen, K.; Ruuskanen, J. Alternative QSAR models for selected estradiol and cytochrome P450 ligands: comparison between classical,

spectroscopic, CoMFA and GRID/GOLPE methods. *SAR QSAR Environ. Res.* **2005,** *16* (6), 555-565.
510. Brown, P. J.; Vannucci, M.; Fearn, T. Bayes model averaging with selection of regressors. *J. R. Statist. Soc. B.* **2002,** *64*, 519-536.
511. Hawkins, D. M.; Subhash, C. B.; Mills, D. Assessing Model Fit by Cross-Validation. *J. Chem. Inf. Comp. Sci.* **2003,** *43* (2), 579-586.
512. Karthikeyan, M.; Glen, R. C.; Bender, A. General melting point prediction based on a diverse compound data set and artificial neural networks. *J. Chem. Inf. Mod.* **2005,** *45* (3), 581-590.
513. Fontaine, F.; Pastor, M.; Zamora, I.; Sanz, F. Anchor-GRIND: Filling the gap between standard 3D QSAR and the GRid-INdependent Descriptors. *J. Med. Chem.* **2005,** *48* (7), 2687-2694.
514. Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comp. Sci.* **2000,** *40* (3), 773-777.
515. Huuskonen, J. J.; Villa, A. E. P.; Tetko, I. V. Prediction of partition coefficient based on atom-type electrotopological state indices. *J. Pharm. Sci.* **1999,** *88* (2), 229-233.
516. Stone, M.; Brooks, R. J. Continuum Regression - Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least-Squares, Partial Least-Squares and Principal Components Regression. *J. R. Statist. Soc. B.* **1990,** *52* (2), 237-269.
517. Helland, I. S. Some theoretical aspects of partial least squares regression. *Chemom. Intell. Lab. Syst.* **2001,** *58* (2), 97-107.
518. Hansch, C. Structure-Activity-Relationships of Chemical Mutagens and Carcinogens. *Sci. Total. Environ.* **1991,** *109*, 17-29.
519. Wold, S. Nonlinear Partial Least-Squares Modeling .2. Spline Inner Relation. *Chemom. Intell. Lab. Syst.* **1992,** *14* (1-3), 71-84.
520. Berglund, A.; Wold, S. INLR, implicit non-linear latent variable regression. *J. Chemom.* **1997,** *11* (2), 141-156.
521. Berglund, A.; Kettaneh, N.; Uppgard, L. L.; Wold, S.; Bendwell, N.; Cameron, D. R. The GIFI approach to non-linear PLS modeling. *J. Chemom.* **2001,** *15* (4), 321-336.
522. Korhonen, S.-P.; Tuppurainen, K.; Laatikainen, R.; Peräkylä, M. Comparing the Performance of FLUFF-BALL to SEAL-CoMFA with a Large Diverse Estrogen Data Set: From Relevant Superpositions to Solid Predictions. *J. Chem. Inf. Mod.* **2005,** *45* (6), 1878-1883.

Kuopio University Publications C. Natural and Environmental Sciences

C 185. Luomala, Eeva-Maria. Photosynthesis, chemical composition and anatomy of Scots pine and Norway spruce needles under elevated atmospheric CO2 concentration and temperature.
2005. 137 p. Acad. Diss.

C 186. Heikkinen, Lasse M. Statistical estimation methods for electrical process tomography.
2005. 147 p. Acad. Diss.

C 187. Riihinen, Kaisu. Phenolic compounds in berries.
2005. 97 p. Acad. Diss.

C 188. Virkutyte, Jurate. Heavy metal bonding and remediation conditions in electrokinetically treated waste medias.
2005. 133 p. Acad. Diss.

C 189. Koistinen, Kaisa. Birch PR-10c: multifunctional binding protein.
2006. 79 p. Acad. Diss.

C 190. Airaksinen, Sanna. Bedding and manure management in horse stables: its effect on stable air quality, paddock hygiene and the compostability and utilization of manure.
2006. 91 p. Acad. Diss.

C 191. Asikainen, Arja. Use of computational tools for rapid sorting and prioritising of organic compounds causing environmental risk with estrogenic and cytochrome P450 activity.
2006. 51 p. Acad. Diss.

C 192. Ålander, Timo. Carbon composition and volatility characteristics of the aerosol particles formed in internal combustion engines.
2006. 54 p. Acad. Diss.

C 193. Molnár, Ferdinand. Structural analysis of the ligand-binding domains of human and mouse CAR, human VDR and human PPARs.
2006. 115 p. Acad. Diss.

C 194. Kasurinen, Anne. Soil-related processes of young silver birch trees grown under elevated CO_2 and O_3 .
2006. 64 p. Acad. Diss.

C 195. Metsärinne, Sirpa. Degradation of Novel and Conventional Complexing Agents.
2006. 138 p. Acad. Diss.

C 196. Heijari, Juha. Seed origin, forest fertilization and chemical elicitor influencing wood characteristics and biotic resistance of Scots pine.
2006. 39 p. Acad. Diss.

C 197. Hakulinen, Mikko. Prediction of density, structure and mechanical properties of trabecular bone using ultrasound and X-ray techniques.
2006. 84 p. Acad. Diss.

C 198. Al Natsheh, Anas. Quantum Mechanics Study of Molecular Clusters Composed of Atmospheric Nucleation Precursors.
2006. 55 p. Acad. Diss.